Geometry of Optimal Design and Limit Theorems

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von Frank Röttger, M.Sc.

geb. am 05.07.1993 in Gronau (Westf.)

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Thomas Kahle
Prof. Dr. Rainer Schwabe
Prof. Henry Wynn, PhD

eingereicht am: 07.01.2020

Verteidigung am: 10.06.2020

Zusammenfassung

Diese Arbeit untersucht das Gebiet der optimalen Versuchsplanung im Sinne der algebraischen Statistik sowie Grenzwertsätze für eine zweiseitige Teststatistik auf Coxeter-Gruppen. Optimale Versuchsplanung für verallgemeinerte lineare Modelle (GLM) ist durch die zunehmende Verbreitung von GLM in den Anwendungen ein wichtiges Thema. Aufgrund ihrer nichtlinearen Struktur gibt es vielfältige Probleme bei der Berechnung von optimalen Versuchsplänen für GLM, insbesondere da die Optimalität in der Regel nur lokal im Parameterraum gegeben ist. Diese Arbeit soll zeigen, wie durch Anwendung von algebraischen und geometrischen Methoden die Komplexität der Berechnung von optimalen Versuchsplänen für GLM verringert werden kann. Die Arbeit steht damit in der Tradition anderer Arbeiten auf dem Gebiet der algebraischen Statistik.

Im ersten Teil der Arbeit werden die notwendige Notation eingeführt und die relevanten mathematischen Grundlagen der statistischen Versuchsplanung, GLM und insbesondere der statistischen Versuchsplanung für GLM erläutert. Nach dieser Einführung beschreiben wir optimale Versuchspläne bezüglich des D-Kriteriums für zwei spezielle GLM. Dabei handelt es sich zunächst um das Bradley–Terry Paarvergleichsmodell, in dem die Präferenz zwischen zwei Alternativen durch eine binäre Zufallsvariable modelliert wird, sodass eine statistische Rangordnung zwischen den Alternativen ersichtlich wird. Wir erhalten eine kombinatorische Beschreibung von D-optimalen Versuchsplänen mit minimalem Träger für das Modell mit beliebig vielen Alternativen und untersuchen die semi-algebraische Geometrie für optimale Versuchspläne für das Modell mit 4 Alternativen. Danach betrachten wir ein spezielles lineares Regressionmodell mit zufälligen Koeffizienten, für welches die Berechnung von optimalen Versuchsplänen durch die Symmetrie und Invarianz des Modells vereinfacht wird. Darauf aufbauend definieren wir die neue Familie der rhombischen Designs und zeigen, dass die D-Optimalität dieser Versuchspläne direkt von der Korrelationsmatrix der zufälligen Koeffizienten abhängt.

Im zweiten Teil der Arbeit präsentieren wir Ergebnisse aus dem Bereich der probabilistischen Kombinatorik. Wir beweisen den zentralen Grenzwertsatz für eine zweiseitige Teststatistik auf Coxeter-Gruppen, also für eine Abbildung, die jedem Gruppenelement und seiner Inverse einen ganzzahligen Wert zuordnet. Die untersuchte Statistik bildet dabei ein Elemtent w aus einer Coxeter-Gruppe W auf die Anzahl der Descents in w plus die Anzahl der Descents in seiner Inversen w^{-1} ab. Wir zeigen, dass für eine zufällige Folge von Gruppenelementen, welche gleichverteilt aus einer Folge von Coxeter-Gruppen gezogen werden, die Statistik genau dann dem zentralen Grenzwertsatz genügt, wenn die Varianz der Stastik gegen unendlich läuft.

Abstract

This thesis investigates the design of experiments from the perspective of algebraic statistics as well as limit theorems for a two-sided test statistic on Coxeter groups. Design of experiments for generalized linear models (GLM) is a recurring topic for practitioners and statisticians, which is of growing importance due to the widespread application of GLM throughout the sciences. Unfortunately, optimal designs for GLM are often hard to obtain and complicated to apply, as optimality can in general only be achieved locally in the parameter space. This thesis aims at showing a path on how to apply tools from algebra and geometry to reduce the complexity of computing experimental designs for GLM in the tradition of previous research conducted in this branch of the field known as algebraic statistics.

In the first part of the thesis, we introduce the notational and mathematical foundations of optimal design, GLM and optimal design for GLM. Based on this introduction, we find designs that are optimal with respect to the special but very important D-criterion for two particular models. The first model we investigate is the Bradley–Terry paired comparison model. In this model, each participant voices a preference for one alternative over another, which reveals a statistical ranking among the alternatives in the experiment. We obtain a combinatorial description of D-optimal designs with minimal support for an arbitrary number of alternatives and study the semi-algebraic geometry of optimality regions for the case with 4 alternatives. Afterwards, we present a special random coefficient model and exploit the symmetry of the model to reduce the complexity of computing optimal designs. We introduce the notion of rhombic designs that suffices to the symmetry of the model and show that the conditions on these designs to be optimal depend directly on the correlation matrix of the random coefficients.

In the second part of the thesis, we present results of a different flavor that belong to the field of probabilistic combinatorics. We show a central limit theorem for a two-sided statistic on Coxeter groups, that is a map from a group element and its inverse to the integers. The statistic we study assigns to an element w of a finite Coxeter group W the number of descents of w plus the number of descents of its inverse w^{-1}. Our main result is that the statistic evaluated at a random sequence of group elements uniformly chosen from Coxeter groups of growing rank satisfies the central limit theorem if and only if the variance of the statistic goes to infinity.

Contents

Acknowledgements

First of all, I want to thank my PhD advisors Thomas Kahle and Rainer Schwabe for their support and guidance, both personally and professionally. Their doors were always open, any question or idea was discussed right away, which created an atmosphere of encouragement. I also want to thank them and my other coauthors Ulrike Graßhoff, Heinz Holling and especially Benjamin Brück for their fruitful collaboration.

I thank my colleagues at the Institute for Mathematical Stochastics for the great time I had there, in particular during lunch, coffee and active sport breaks. Here, special thanks go to Kerstin Altenkirch for taking very good care for all of us. I also thank everybody at the Institute for Algebra and Geometry for immediately including me into their institute.

My position was part of and funded by the research training group MathCoRe. I would like to express special gratitude to all of the PI's and fellows, especially my mentor Eliana Duarte and the speakers Sebastian Sager and Volker Kaibel.

I thank my family and friends for their everlasting support and friendship. My deepest thanks go to Melanie for her love and for sharing her life with me.

1. Introduction

In the introduction, we give a short overview over the history and current situation of design of experiments for generalized linear models and its connection with algebraic statistics, as well as a draft of the discussed topics of this thesis.

The thought of statistical planning develops naturally in any statistician conducting an experiment. Early systematic investigations on how to plan experiments developed into the field of *design of experiments*, which is also known as *optimal design* or *experimental design* in the literature. Various authors contributed to the theory in the first half of the 20th century, for example Kirstine Smith in 1918 [Smi18], before Ronald Fisher published his monograph *The design of experiments* in 1935 [Fis35], which is predominantly named as the foundational book of the field. Fisher worked from 1919-1933 at Rothamsted Experimental Station in one of the oldest agricultural research facilities in the world, where he investigated design of experiments for crop studies. Until today, agricultural experiments, together with medical and pharmaceutical studies as well as trials in psychology or marketing, are among the most important applications for the theory.

The basic concept of optimal design is the following: A practitioner is planning an experiment and needs to decide how to collect the data in the experiment with respect to a chosen statistical model, so how many observations to draw at each experimental setting. The way the observations are distributed among the possible experimental settings is called an *experimental design*. Obviously, the experimenter is interested in maximizing the statistical *information* that can be obtained by the experiment that is conducted according to the selected design. The information of some experiment is not an abstract concept, but was formalized by Fisher as the variance of the score-function, which is the derivative of the likelihood function. Consequently, the information is also known as *Fisher information*. The definition is justified by the fact that the covariance matrix of suitable estimators like the maximum likelihood estimator coincides with the inverse of the Fisher information. Now, the information of the experiment varies among the possible experimental designs. In an experiment with only one parameter, the optimal design that maximizes the information would therefore be the one with the highest information, but for multiparametric problems, the Fisher information is a matrix (cf. Definition 2.12). To define a maximum, the typical approach of the theory of optimal design is to maximize some functional which is mapping the information matrix to the real line instead. Such a functional is called an *optimality criterion*. Two of the standard examples are listed in Section 2.8.

Following the work of Fisher, the theory was pushed forward in the second half of the 20th century by researchers like Rao, Elfving, Chernoff, Kiefer, Wolfowitz and many others. Kiefer and Wolfowitz proved the first equivalence theorem in 1960 [KW60]. It shows

the equivalence of D-optimality (cf. Section 2.8.1) and G-optimality (cf. Section 2.8.2) in linear models. In this thesis, we restrict ourselves to the case of D-optimality, which means that a design is denoted as optimal when it maximizes the logarithm of the determinant of the information matrix. Because of the asymptotic relation between the inverse information matrix and the covariance of the maximum likelihood estimator, the D-criterion corresponds to minimizing the volume of the confidence ellipsoid. As the set of information matrices is convex and the D-criterion is concave, the process to find an optimal design with respect to the D-criterion is in principle a convex optimization problem. This enabled Kiefer and Wolfowitz to apply the toolset of convex optimization theory to experimental design. Further developments of the theory lead to the general equivalence theorem proven by Whittle in [Whi73], which shows a duality result for a general optimality criterion under certain assumptions. We introduce the generalized Kiefer–Wolfowitz equivalence theorem, adapted to our application, in Section 2.10. Among the standard references for design of experiments are the books of Fedorov [Fed72], Pukelsheim [Puk93] and Silvey [Sil80], to which we will refer frequently throughout the thesis.

In the last decades, a new branch of statistics developed: *Algebraic statistics*. The leitmotif of this field is the application of techniques from commutative algebra, algebraic geometry and computer algebra to problems in statistics. One of the first fields in statistics where this approach was picked up is design of experiments. Consequently, the term *algebraic statistics* was coined by Riccomagno, Pistone and Wynn in [PRW00]. In their book they study, aside of other questions of algebraic statistics, the question of identifiability of regression models via experimental designs. A regression model is identifiable by an experimental design, when the corresponding information matrix is non-singular. In particular, Riccomagno et al. investigate the identifiability of polynomial regression models with techniques from computer algebra. For this question, they do not consider the actual proportion of observations at each *design point*, but only the support (thus the design points) of a design. This means that in their definition, a design is a finite set of points in $\mathbb{K}^d$. The design corresponds to the set of all polynomials in $\mathbb{K}[x_1, \ldots, x_d]$ that evaluate to zero in all of the design points. Such a set of polynomials is called a (polynomial) ideal. Each ideal can be represented by a finite set of polynomials which is referred to as a basis. A special type of basis that exhibits useful properties for computational approaches is a Gröbner basis. Pistone et al. show how to derive whether a polynomial regression model is identifiable by a chosen design from its corresponding Gröbner basis [PRW00, Theorem 27].

An overview over the history of algebraic statistics is [Ric09]. According to this article, the first publications in the field of algebraic statistics are the articles of Pistone–Wynn [PW96] and Diaconis–Sturmfels [DS98], although some algebraic statisticians consider the article of Pearson from 1894 [Pea94] as the origin of algebraic statistics. Pearson showed that the empirical distribution of the ratio between the forehead- and body-length of crabs in the bay of Naples can be explained with a Gaussian mixture. To do this, he had to use the method of moments to compute a polynomial of degree 9, see [ARS17] for a detailed overview. The article of Pistone–Wynn [PW96] studies

the identifiability of models given a particular design, see also [PRW00], as outlined above. Diaconis and Sturmfels introduce a new algorithm for Markov chain sampling that relies on computations over polynomial rings and therefore requires methods from computational algebra. Today, among the standard references in algebraic statistics are the books of Sullivant [Sul18] and the lecture notes of Drton, Sturmfels and Sullivant that arose from an Oberwolfach seminar [DSS09]. In [Sul18] there is a survey of results from design-flavored algebraic statistics in Chapter 12. Drton et al. [DSS09] do not discuss design of experiments. Further recent references for algebraic statistics in experimental design are [PWZ17, RKR16, MSW13, AHT12, BMO+10, GRRW10, BLM+07].

Generalized linear models (GLM) are a generalization of linear regression models. As they are highly relevant in modern statistics, there is a significant interest in optimal designs for such models. For GLM, the Fisher information depends non-linearly on the unknown parameter of interest. Therefore, D-optimal designs are only locally optimal, which means that the optimality of a design depends on the parameter. Furthermore, through the non-linearity, the computational complexity to find an optimal design increases quickly in the dimension. We introduce GLM in Section 2.2 and exhibit design of experiments for GLM in Section 2.6. Local optimality is explained in Section 2.9. We study optimal designs for a specific GLM in Chapter 3 in form of the Bradley–Terry paired comparison model, which is a logistic regression model with a specific structure of the experimental settings. An introduction to Chapter 3 follows in Section 1.2.

As we exhibited before, the problem to find a D-optimal design translates into a convex optimization problem. In Section 2.11, we formalize this approach for GLM. We show that the optimality region of a locally D-optimal design in the parameter space for GLM under the assumption of a discrete design region as well as polynomial entries in the regression function is given as a semi-algebraic set, that is a set defined by polynomial inequalities and equations (see Definition 2.36). If a design has a positive weight on each design point in the design region, the corresponding optimality region in the parameter space is the non-negative real part of an affine variety, so the zero set of a finite collection of polynomials (see Definition 2.35). This follows from the fact that if we choose the correct variables, the entries of the information matrix of a design for a GLM are polynomials. Via Cramer's rule and the adjugate matrix, the polynomial structure in the optimization problem follows (see (2.11.1)).

This thesis is based on four articles: [KRS19, GHRS20, Röt20] and [BR19]. The structure of the thesis is as follows: Chapter 2 introduces the theory of generalized linear models, optimal design and basics from algebra that are necessary to study the geometry of optimal design. In Chapter 3 we exhibit the geometry of the Bradley–Terry paired comparison model, which is summarized in Section 1.2. Chapter 3 was published as [KRS19]. Afterwards, we discuss local optimality for a special type of linear regression with random coefficients in Chapter 4, which was published as [GHRS20]. An introduction to this chapter is Section 1.3. In the second part of the thesis, we present the results of [Röt20] and [BR19], which show the central limit theorem for a two-sided statistic on Coxeter groups, in Chapter 5. An introduction is given in Section 1.4.

1.1. Design of experiments in data science

In this section, we give a short introduction to the connection of experimental design to recent topics in statistics and their applications.

A growing topic in the intersection of statistics and computer science is a field that is, in general, called "big data". The name is somewhat self-explanatory: Through modern technology, data is collected and analyzed automatically on a large scale. This leads to the problem that standard statistical methods are not of reasonable computation time due to the sheer size of the data set. This is one of the problems that sparks the interest in the field of machine learning as a tool to deal with the computational and statistical challenges that go along with large data sets.

Large data sets are often obtained as a byproduct or at least without direct control of the covariates in the data acquisition. Therefore, the importance of design of experiments for these problems is not directly visible. It is revealed while handling the data itself. An example is presented in the paper of Drovandi et al. [DHM+17], where they discuss the application of design of experiments for model selection in large data sets. Their idea is to choose a subsample with respect to some sampling design and choose the model from this subsample. This can also be applied sequentially with a sequential design if there is a "continuous" inflow of data.

A common problem of passively obtained large data sets is bias, as potentially hidden confounders may influence the observations. Pesce et al. [PRW19] discuss this problem and suggest to embed the model into a larger model with an added bias term that represents the hidden confounders. By the means of the randomization the bias can be modeled as a block effect. Pesce et al. propose a game-theoretic approach to minimize the covariance of the maximum likelihood estimator for the parameter of the model while balancing the potential bias.

Another field with connections to optimal design is the study of bandit problems, which are a standard topic in machine learning. Already in 1960, Vogel [Vog60] introduced sequential allocation designs for two-armed bandit experiments. A recent contribution to this problem is the article of Zhang and Lee [ZL10], where they apply computational methods and empirical priors to optimize sequential allocation designs for bandit models.

The field of design of experiments studies statistical models via the corresponding Fisher information. A related field that generalizes the investigation of statistical models with a perspective from information theory and differential geometry is known as information geometry. A recent textbook is [AJLS17] and a short introduction is [Pis19]. A prime example discussed in [Pis19] is the probability simplex (cf. Definition 2.25) for discrete statistical models. For example, say that we study an experiment with two binary random variables X_1, X_2. This is parameterized by

$$p_{11} = \mathbb{P}(X_1 = 1, X_2 = 1), \qquad p_{10} = \mathbb{P}(X_1 = 1, X_2 = 0), \qquad p_{01} = \mathbb{P}(X_1 = 0, X_2 = 1).$$

with $p_{00} = \mathbb{P}(X_1 = 0, X_2 = 0) = 1 - p_{11} - p_{10} - p_{01}$. The corresponding probability simplex is a convex set in $\mathbb{R}^3$ spanned by the origin and the three unit vectors. Now,

when we impose constraints on the model, we see that these constraints correspond to subsets of the probability simplex. For example, when we require X_1 and X_2 to be independent, the resulting model corresponds to a surface in the probability simplex defined by

$$p_{00}p_{11} = p_{10}p_{01}.$$

This is a twisted square inside the simplex where the corners coincide with the vertices.

A formalized approach to the study of statistical models is the definition of a statistical manifold. A statistical manifold is a Riemannian manifold where each point corresponds to a probability distribution. A Riemannian metric that links information geometry with design of experiments is the Fisher metric, introduced by Rao in 1945 [Rao45]. The Fisher metric defined on a smooth statistical manifold is, if seen as a matrix, the Fisher information matrix (see Definition 2.12 and [AJLS17, Def. 2.1]). For example, the interior of the probability simplex is a smooth statistical manifold. For a study of the boundary of statistical manifolds like the probability simplex, see [Kah10].

In the example above, the joint density of X_1 and X_2 parameterizes the probability simplex. This makes it possible to obtain general formulas for the information matrix and its properties, see [Pis19, Prop. 13]. For example, the inverse information matrix is zero only on the vertices of the probability simplex. Furthermore, the determinant of the inverse of the information matrix is computed, which shows that it equals zero on the borders of the probability simplex and is positive on the inside. This approach generalizes from the submanifold that is the interior of the probability simplex to statistical manifolds, see [AJLS17].

In this thesis, we study a geometric perspective on topics in optimal design for GLM, which is a contribution to integrate this particular aspect into data science. For example, learning deep neural networks can be interpreted as the iterated application of GLM. The analysis of large data sets creates the necessity for automated approaches in the theory. Further collaboration between algebra, geometry and statistics may provide new methods for such problems in modern statistics.

1.2. Introduction to the chapter on the Bradley–Terry paired comparison model

Consider an experiment to evaluate the preferences among m alternatives in which participants choose one preferred item when presented two of the alternatives. Thus, the choices of the participants are binary replies. This setting can be studied with the Bradley–Terry paired comparison model, which was introduced in [BT52] to analyze taste testing results for pork depending on different feeding patterns. The model has proven popular in different areas of statistics. Hastie and Tibshirani [HT98] developed a coupling model similar to the Bradley–Terry model to study class probabilities for pairs of classes. Simons and Yao [SY99] discussed the asymptotics of the model when the number of potential alternatives tends to infinity. Algorithms for Bradley–Terry

models are discussed, for example, in [Hun04], and asymptotics of algorithms, for example, in [DMJ13]. Besides marketing or transportation, another popular application area for the Bradley–Terry model is the world of professional sports such as American football, car racing, matching in tournaments, card games or strategies for sport bets, see [CMP07, GRF03, BMS04, KKT06]. The Bradley–Terry model is part of a broader class of models that describe statistical rankings. Specifically, it arises from marginalization of the Plackett–Luce model, see [SW12].

We are interested in optimal designs for the Bradley–Terry model, that is, a scheme how to assign a fixed number of measurements to different experimental settings, such that the experiment is most informative. Optimal experimental designs for the Bradley–Terry model were first investigated by Torsney [Tor04], who gave an algorithmic approach to fit the parameters of the model. Graßhoff and Schwabe [GS08] completely analyzed the case of three alternatives. They gave symbolic solutions for the design problem depending on the parameters and described the optimality areas of these design classes in the parameter space. We extend the results of Graßhoff and Schwabe in two directions. We discuss the case of four alternatives in detail and characterize optimal saturated designs, that is designs with minimal support size, for an arbitrary number of alternatives. Section 3.1 gives the general setup. Section 3.4 contains an almost complete analysis of the case of 4 alternatives. Only one very challenging polynomial inequality system remains open (Problem 3.10). In Sections 3.2 and 3.3 we discuss saturated optimal designs for an arbitrary number of alternatives. Our main result is an easy, combinatorial polynomial inequality description of regions in the parameter space, where a given saturated design is optimal. We include the information for which designs these region of optimality are empty (Theorem 3.6).

Theorem 1.1 (Simplified version of Theorem 3.6)**.** *For the Bradley–Terry paired comparison model, a design with minimal support is D-optimal if and only if the graph, where the nodes correspond to the alternatives and the edges to the comparisons that appear in the design, is a path.*

Polynomial inequality constraints in experimental design are a recurring topic. In [KOS16], Kahle et al. discuss the optimality of saturated designs for the Rasch Poisson counts model, which is an example for Poisson regression. They explain for a special case, where the information matrix polytope (cf. Definition 2.22) equals the correlation polytope, how to relax the problem to find a D-optimal design by passing over to the linear matrix inequality relaxation of the information matrix polytope. This means that instead of optimizing over the information matrix polytope, one optimizes over the corresponding spectahedron that arises as an intersection of the cone of positive semidefinite matrices and the affine space spanned by the information matrix polytope [KOS16, Section 4]. As Poisson regression is a GLM, their results can be discussed in the setting of Section 2.6.

Knowledge about the optimality regions can be very helpful in designing experiments. For example, a pilot study could reveal that the estimate of the parameters are all within one region of optimality. In such a situation it is then clear which design to use. See [DMP+04] for a general class of models where local optimality is studied. In the

discussion in Section 3.5 we compare the efficiency of our tailored designs versus uniform designs as the parameters grow in magnitude. Furthermore, we outline open problems and further research directions.

1.3. Introduction to the chapter on optimality regions in multiple regression with correlated random coefficients

Hierarchical regression models with random coefficients enjoy growing importance in biological and psychological applications, whenever there is a variation with respect to the observed subjects. In a random coefficient model, one assumes that the parameters are random variables instead of deterministic values. This is applied to for example capture individual differences among the participants of an experiment that cannot be explained by a fixed parameter. We often cannot expect the random coefficients to be uncorrelated. Hence, we assume that the random coefficients are e.g. normally distributed with a population mean and a non-diagonal covariance matrix. A special model that will be the topic of this chapter are random effects models for linear regression with single responses. This means that we obtain only one observation per unit. This particular model was motivated by Freund and Holling in [FH08] and Patan and Bogacka in [PB07]. One wants to find optimal experimental designs for these models with respect to some optimality criterion. Graßhoff et al. determined D-optimal designs that maximize the determinant of the corresponding information matrix, for a couple of different covariance structures in [GHS09] and [GDHS12]. They found that in contrast to fixed effects models for multiple linear regression optimal settings may, surprisingly, occur in the interior of the design region under certain conditions on the covariance structure of the random coefficients. We assume that the random coefficients are distributed with equal variances and are equi-correlated as well as uncorrelated with the random coefficients of the linear effects. Now, we investigate conditions on the covariance structure to discriminate situations in which optimal designs are completely supported on the extreme points of the design region as in fixed effects models and situations in which optimal designs must have additional support points in the interior. This discrimination is done for the special class of *rhombic* designs, which are invariant with respect to permutations of the regressors and simultaneous sign change. We introduce rhombic designs in Section 4.2. Section 4.3 studies via the Kiefer–Wolfowitz equivalence theorem (Theorem 2.33) for which parameter regions a D-optimal rhombic design has interior support points. Said parameter regions are described by semi-algebraic sets. Furthermore, we discuss how the optimality depends on the covariance structure. Additionally, we show that for the assumed covariance structure of the random coefficients, the D-optimality of designs with interior support points translates to a simple matrix equation for the information. Let $D = \operatorname{diag}(d_0, D_1)$ be the $(K+1) \times (K+1)$-dimensional covariance matrix of the random coefficients, so that $d_0 > 0$ and D_1 is a positive semidefinite $K \times K$-dimensional *completely symmetric matrix* (cf. Eq. (4.2.3)). Furthermore, let $M(\xi)$ denote the infor-

mation matrix of some experimental design ξ.

Theorem 1.2 (Simplified version of Theorem 4.5)**.** *A rhombic design ξ^* that is supported only on the vertices of the hypercube is D-optimal if and only if $D-(K+1)M(\xi^*)^{-1}$ is a diagonal matrix with vanishing trace and non-negative first entry.*

Theorem 1.3 (Simplified version of Theorem 4.8)**.** *A design ξ^* with a design point in the interior of the hypercube is D-optimal if and only if $(K+1)M(\xi^*) = D^{-1}$.*

We show as a consequence of the results in Section 4.3 that the distinction, whether a D-optimal rhombic design requires interior support points or not, can be made by evaluating a polynomial that only depends on the covariance matrix of the random coefficients.

Corollary 1.4 (Simplified version of Corollary 4.6 and Corollary 4.9)**.** *A rhombic design supported (not) only on the vertices of the hypercube can only be D-optimal when the first diagonal entry of D^{-1} is (smaller) larger or equal to the second diagonal entry of D^{-1}.*

Based on these results, we are able to compute optimal designs and their optimality regions explicitly for small to moderate dimensions in Section 4.4 and we conjecture results for arbitrary dimensions in Section 4.5.

1.4. Introduction to the chapter on the central limit theorem for a two-sided statistic on Coxeter groups

Statistical and probabilistic methods are powerful tools for the investigation of combinatorial and algebraic objects. They reveal deeply rooted connections between those fields. Of greatest importance in probabilistic asymptotics is the central limit theorem (CLT), that is the convergence in distribution of a sequence of random variables, normalized by its mean and its standard deviation, towards the standard Gaussian. The main result of the chapter is an equivalent formulation of the central limit theorem for a sequence of random variables that arises from a statistic on sequences of finite Coxeter groups. Chatterjee–Diaconis [CD17] denominate the statistical relevance of the CLT for the two-sided descent statistic to be for permutation tests. In a permutation test, we are interested in certificates for "non-randomness" of a sample. The research of Chatterjee and Diaconis regarding the two-sided descent statistic was motivated by defining a metric from descents to use as test statistic. As the descent statistic is not symmetric, the two-sided descent statistic is preferred (although it also does not define a metric, cf. [CD17]).

Finite Coxeter groups or reflection groups are a recurring topic in (algebraic) statistics and probabilistic combinatorics. Beside the research on asymptotics, the importance of reflection groups is well known in many fields of statistics. In the last decades, this extended from the symmetric group to other reflection groups, especially to the infinite

irreducible families such as the signed permutation groups. A standard example, as in this thesis, is the theory of design of experiments, where invariance and symmetry considerations are a useful tool (see e.g. Chapters 3 and 4). This goes back at least to Kiefer [Kie74]. In algebraic statistics, recent examples are the paper of Francis et al. [FSW17], where they discuss the construction of exact confidence nets via finite reflection groups and the work of Boege et al. [BDKS19, Section 3], where the geometry of gaussoids reveals a symmetry imposed by the signed permutation group.

In the symmetric group $\mathrm{Sym}(n)$, which is the Coxeter group of type $\mathtt{A}_{n-1}$, the descent statistic is defined as follows: Write the elements of $\mathrm{Sym}(n)$ in their one-line notation. Then, the number of descents $\mathrm{des}(\pi)$ of an element $\pi \in \mathrm{Sym}(n)$ is the number of positions in the corresponding one-line notation where an entry is larger than its successor. This concept generalizes to arbitrary finite Coxeter groups, the necessary definitions are presented in Section 5.3.

Fixing such a Coxeter group W, choosing an element of W uniformly at random and evaluating the descent statistic gives rise to a random variable D_W. Kahle and Stump recently showed that for sequences $(W_n)_n$ of finite Coxeter groups of growing rank, the sequence D_{W_n} satisfies the CLT if and only if its variance tends to infinity, see [KS20]. They asked [KS20, Problem 6.10] whether for the random variable T_W associated to the statistic $t(w) \coloneqq \mathrm{des}(w) + \mathrm{des}(w^{-1})$, a similar statement holds true. The statistic t also has a geometric interpretation in terms of a two-sided analogue of the Coxeter complex introduced by Petersen [Pet18], for details see [BR19, Appendix A].

For the case where $W_n = \mathrm{Sym}(n+1)$, the irreducible Coxeter group of type $\mathtt{A}_n$, Vatutin [Vat96] showed that T_W satisfies the CLT. This was later, with different methods, reproven by Chatterjee–Diaconis [CD17] and Özdemir [Ö19]. To obtain a similar result for the group $W_n = \mathtt{B}_n$, so for signed permutations, we follow the approach of Chatterjee–Diaconis. They generated a random permutation and its inverse from a vector of random variables that are uniformly distributed on the square (cf. Section 5.7) and applied the method of interaction graphs, which is shortly introduced in Section 5.5. We modified this approach by adding a random sign in the right way (see Section 5.8). The result in Section 5.9 is the CLT for T_W where $W_n = \mathtt{B}_n$. Consequently, in Section 5.10 we derive the CLT for T_W where $W_n = \mathtt{D}_n$ by exploiting the similarities between $\mathtt{B}_n$ and $\mathtt{D}_n$.

Theorem 1.5 (Simplified version of Theorems 5.11 and 5.12)**.** *Let $W_n = \mathtt{B}_n$ or $W_n = \mathtt{D}_n$. Then, T_W satisfies the CLT.*

Section 5.15.1 generalizes the above result to a broader class of statistics. Section 5.15.2 shows via the Cramér-Wold device that Theorem 1.5 implies a two-dimensional CLT for the statistic $(\mathrm{des}(w), \mathrm{des}(w^{-1}))$.

Our main result is a positive answer to the question of Kahle–Stump under an additional hypothesis on the sequence of Coxeter groups:

Theorem 1.6 (Simplified version of Theorem 5.37)**.** *Let $(W_n)_n$ be a well-behaved sequence of finite Coxeter groups such that $\mathrm{rk}(W_n) \to \infty$ and let T_n be the random variable associated to the statistic t on W_n. Then the following are equivalent:*

1. *$(T_n)_n$ satisfies the CLT.*

2. *The variance of T_n goes to infinity.*

3. *The maximal size $\max_n$ of a dihedral parabolic subgroup in W_n does not grow too fast (This is in particular the case if $\max_n$ is bounded).*

We give a precise statement of the result as Theorem 5.37 in Section 5.14 but would like to remark that we were not able to construct a sequence of Coxeter groups that is not well-behaved in the above sense. After posting our results to the arXiv, V. Féray [Fé20] provided a proof for Theorem 5.37 that does not require the sequence to be well-behaved, see also Remark 5.39.

In order to obtain the result above, we take an approach similar to the one used by Kahle–Stump [KS20] for the descent statistic. In particular, this involves an application of Lindeberg's theorem for triangular arrays. There is however a major difference between their approach and ours: The generating function of the descent statistic is given by the Eulerian polynomial which factors over the reals and has only negative roots, see [Bre94] and [SV15]. Kahle and Stump heavily used this in order to deduce their result. In contrast to that, the generating function of the statistic t is the two-sided Eulerian polynomial as studied e.g. in [CRS66], [Pet13] and [Vis13]. It does not have such a nice factorization, even in the setting of symmetric groups. In order to resolve the additional difficulties arising from this, we are led to compute higher moments of the random variables T_W in Section 5.11. For this, we use and generalize the work of Özdemir [Ö19]. A list of moments that we computed is in Appendix A. We use the results for the higher moments to apply the Lindeberg theorem for triangular arrays in Section 5.13. Together with the result on a CLT for weighted sums of converging sequences presented in Section 5.12, we obtain the main result. The chapter ends with a discussion of possible future research directions and a generalization of Theorem 1.5 as well as a two-dimensional CLT via the Theorem of Cramér–Wold.

2. Generalized Linear Models and Optimal Design Theory

This chapter introduces definitions and notations for exponential families and generalized linear models following the textbooks of Alsmeyer [Als09] and Shao [Sha03], and the publication of Fahrmeir and Kaufmann [FK85]. Afterwards, we establish the concepts of optimal design of experiments as described in [Sil80] while focusing on generalized linear models and outline a geometric perspective on the corresponding optimization problem.

2.1. Natural Exponential Families

We begin with the basic concept of a statistical experiment:

Definition 2.1. A *statistical experiment* is a triple

$$(\mathcal{S}, \mathcal{F}, (W_\theta)_{\theta\in\Theta}),$$

consisting of an non-empty set $\mathcal{S}$, a σ-algebra $\mathcal{F}$ on $\mathcal{S}$ and a family $(W_\theta)_{\theta\in\Theta}$ of probability measures on $(\mathcal{S}, \mathcal{F})$, which is parameterized by elements of a parameter space Θ.

Before we introduce exponential families, we remind the reader of the notions of dominated measures and dominated statistical experiments:

Definition 2.2. Let $\tilde{\nu}$ and ν be two measures defined on the same measurable space. Then we say that $\tilde{\nu}$ is *dominated* by ν, if $\tilde{\nu}(A) = 0$ for every set A where $\nu(A) = 0$.

Definition 2.3. Let $(\mathcal{S}, \mathcal{F}, (W_\theta)_{\theta\in\Theta})$ be a statistical experiment. We say that the experiment is *dominated*, if there exists a σ-finite measure ν on $(\mathcal{S}, \mathcal{F})$, so that W_θ is dominated by ν for all $\theta \in \Theta$.

By the Theorem of Radon-Nikodym(see for example [Bil95]), it follows that if a statistical experiment is dominated there exist ν-densities

$$p_\theta := \frac{dW_\theta}{d\nu}$$

for all $\theta \in \Theta$. Now, we are able to define exponential families:

Definition 2.4 (One-parametric exponential family)**.** Let $(\mathcal{S}, \mathcal{F}, (W_\theta)_{\theta\in\Theta})$ be a dominated statistical experiment with dominating measure ν. The family $(W_\theta)_{\theta\in\Theta}$ of probability measures is a one-parametric *exponential family*, if the ν-densities $p_\theta = \frac{dW_\theta}{d\nu}$ are of the following form: There are maps $C : \Theta \to \mathbb{R}$, $Q : \Theta \to \mathbb{R}$ as well as measurable functions $h : \mathcal{S} \to \mathbb{R}$, $T : \mathcal{S} \to \mathbb{R}$ so that for every $\theta \in \Theta$ it holds that

$$p_\theta = C(\theta)h \exp(Q(\theta)T), \ \nu - \text{a.s.} \tag{2.1.1}$$

In applications, it is often the case that we are given a family $(W_\theta)_{\theta\in\Theta}$ via a density p_θ of the form described in (2.1.1) and then confirm that it is dominated by a σ-finite measure. Usually, this measure is either the counting measure or the Lebesgue measure.

Example 2.5. Many of the standard distributions are exponential families, we state a few one-parametric examples below:

- Bernoulli distribution with $Q(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, $C(\theta) = 1-\theta$, $T(y) = y$ and $h(y) = 1$ with the counting measure as the dominating measure and $\Theta = [0, 1]$,
- Poisson distribution with $Q(\theta) = \log(\theta)$, $C(\theta) = e^{-\theta}$, $T(y) = y$ and $h(y) = \frac{1}{y!}$ with the counting measure as the dominating measure and $\Theta = \mathbb{R}_{>0}$,
- Exponential distribution with $Q(\theta) = -\theta$, $C(\theta) = \theta$, $T(y) = y$ and $h(y) = 1$ with the Lebesgue measure as the dominating measure and $\Theta = \mathbb{R}_{>0}$,
- Normal distribution with mean θ, assuming known variance σ^2, with $Q(\theta) = \frac{\theta}{\sigma^2}$, $C(\theta) = e^{-\frac{\theta^2}{2\sigma^2}}$, $T(y) = y$ and $h(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}}$ with the Lebesgue measure as the dominating measure and $\Theta = \mathbb{R}$.

We see from the structure of the density in Eq. (2.1.1) that $C(\theta)$ is the scaling factor which guarantees that the density p_θ integrates to one with respect to ν. Therefore we write

$$C(\theta) = \left(\int h(y) \exp\left(Q(\theta)T(y)\right) \nu(dy) \right)^{-1}.$$

This implies that p_θ only depends on θ via $Q(\theta)$, which allows us to reparameterize the exponential family in terms of $\zeta = Q(\theta)$:

Definition 2.6 (Natural Exponential Family)**.** An one-parametric exponential family with a ν-density of the form

$$p_\zeta = C(\zeta)h \exp\left(\zeta\, T\right)$$

with $\zeta = Q(\theta)$ is called a *natural exponential family* with *natural parameter* ζ. Let

$$\mathcal{Z} := \{\zeta \in \mathbb{R} \mid 0 < \int h \exp\left(\zeta T\right) d\nu < \infty\}$$

be the *natural parameter space*.

$\mathcal{Z}$ is a convex subset of $\mathbb{R}$ that contains the set $Q(\Theta)$. For many models, it holds that $Q(\Theta) = \mathcal{Z} = \mathbb{R}$. Another analytical simplification of exponential families is obtained by passing over to the pushforward measure $W_\theta^T := W_\theta \circ T^{-1}$, which is a probability measure on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Theorem 2.7. *Let $(W_\theta)_{\theta\in\Theta}$ be an exponential family in Q and T. Then, the pushforward measure $(W_\theta^T)_{\theta\in\Theta}$ is an exponential family in Q and the identity.*

The Theorem enables us to study natural exponential families with T as the identity. Let Y be a random variable with ν-density $p_\theta(y) = C(\theta)h(y)\exp(\theta y)$ for $\theta \in \Theta$. Assuming that Θ has interior points, all moments of Y exist for $\theta \in \Theta^\circ$. Additionally assuming $C(\theta) > 0$, we define

$$\mu(\theta) := \mathbb{E}_\theta(Y) = \frac{\partial}{\partial\theta}(-\log(C(\theta))), \qquad \Sigma(\theta) := \mathrm{Cov}_\theta(Y) = \frac{\partial^2}{\partial\theta^2}(-\log(C(\theta))).$$

To visualize this, we exemplify the case of a one-parametric exponential family dominated by the Lebesgue measure. From [Leh86, Theorem 9 in Section 2.7] it follows that the integration with respect to y and the derivative with respect to θ commute. Therefore, it holds that

$$\begin{aligned}
\mathbb{E}_\theta(Y) &= \int y\, C(\theta)h(y)\exp(\theta y)\, dy \\
&= \int C(\theta)h(y)\frac{\partial}{\partial\theta}\exp(\theta y)\, dy \\
&= \int \frac{\partial}{\partial\theta}\left(C(\theta)h(y)\exp(\theta y)\right) - \left(\frac{\partial}{\partial\theta}C(\theta)\right)h(y)\exp(\theta y)\, dy \\
&= -\int \left(\frac{\partial}{\partial\theta}C(\theta)\right)h(y)\exp(\theta y)\, dy \\
&= -\frac{\frac{\partial}{\partial\theta}C(\theta)}{C(\theta)} = -\frac{\partial}{\partial\theta}\log(C(\theta)).
\end{aligned}$$

It is common to use the convenient notation $b(\theta) := -\log(C(\theta))$, so that

$$p_\theta(y) = h(y)\exp\left(\theta^T y - b(\theta)\right).$$

Let Θ° denote the interior of the parameter space Θ. The positive definiteness of $\Sigma(\theta)$ for $\theta \in \Theta^\circ$ implies that $\mu(\theta)$ is injective on Θ° [FK85]. Therefore, a value $\mu(\theta)$ only corresponds to at most one parameter value $\theta \in \Theta^\circ$.

2.2. Generalized linear models

With the definition of a natural exponential family, we can introduce generalized linear models. From now on, θ denotes the natural parameter of a one-parametric exponential family. We will restrict ourselves to the case of one-parametric exponential families.

Definition 2.8 ([FK85, Section 2.1])**.** A generalized linear model is defined as follows: Let $Y_i,\ 1 \leq i \leq n$ be independent random variables that are distributed according to a natural one-parametric exponential family with density

$$p_\theta(y) = h(y) \exp\left(\theta y - b(\theta)\right).$$

Let $f : \mathcal{X} \to \mathbb{R}^p$ be a p-dimensional regression vector. We say that $f(x_i)$ with $x_i \in \mathcal{X}$ relates to Y_i via a linear combination with a parameter $\beta \in B \subseteq \mathbb{R}^p$ such that $\gamma_i = f(x_i)^T\beta$. The linear predictor γ_i connects to the mean $\mu(\theta)$ of Y_i via the injective link function

$$g : \mu(\Theta^\circ) \to \mathbb{R},$$

such that $\gamma_i = g(\mu(\theta))$.

Instead of using the link function $g(\gamma)$, Fahrmeir and Kaufmann claim that it is more convenient for theoretical purposes to introduce an injective function $u = (g \circ \mu)^{-1}$. This implies that $\theta = u(f(x)^T\beta)$. Of special importance are *natural link functions*:

Definition 2.9 (Natural link functions)**.** The link function $g = \mu^{-1}$ is called a natural link function. This implies $u = \mathrm{id}$ and we obtain a linear regression model $\theta = f(x)^T\beta$ for the natural parameter.

A classical example for a generalized linear model is logistic regression:

Example 2.10 (Logistic Regression for binary responses)**.** The model consists of binary random variables $Y_1, \ldots, Y_n$ with

$$\mathbb{E}_{p_i}(Y_i) = \mathbb{P}(Y_i = 1) = p_i, \qquad \mathbb{P}(Y_i = 0) = 1 - p_i. \tag{2.2.1}$$

Therefore, the distribution of Y_i with respect to the counting measure is a natural exponential family with natural parameter $\theta = Q(p) = \log\left(\frac{p}{1-p}\right)$, $h(y) = 1$, $T(y) = y$, $b(\theta) = \log(1 + e^\theta)$ and $\Theta^\circ = \mathbb{R}$. This implies that $\mu(\theta) = \frac{\partial}{\partial\theta} b(\theta) = \frac{e^\theta}{1+e^\theta}$. Now, the natural link function $g = \mu^{-1}$ implies

$$\theta = f(x)^T\beta,$$

so a linear regression for the natural parameter θ. This is equivalent to

$$\log\left(\frac{p}{1-p}\right) = f(x)^T\beta \quad \Leftrightarrow \quad p = \mu(\theta) = \frac{e^{f(x)^T\beta}}{1 + e^{f(x)^T\beta}}.$$

The function $g(p) = \log\left(\frac{p}{1-p}\right)$ is also known as the logit link function. The terms $\frac{p}{1-p}$ are sometimes denoted as odds and consequently $\log\left(\frac{p}{1-p}\right)$ as log-odds.

2.3. Regularity assumptions and the information matrix

Let $f : \mathcal{X} \to \mathbb{R}^p$ be a regression vector. Before we are able to introduce the information matrix for GLM, we need to make certain regularity assumptions. Let β_0 denote the true but unknown parameter, where β is any parameter in some $B \subseteq \mathbb{R}^p$. Usually, $B = \mathbb{R}^p$. We assume the following:

1. B is open and, for natural link functions, convex.
2. $f(x)^T\beta \in g(\mu(\Theta^\circ))$ for all $\beta \in B$.
3. u is twice continuously differentiable.
4. Let $(x_i)_i$ be a sequence on $\mathcal{X}$. The matrix $\sum_{i=1}^n f(x_i)f(x_i)^T$ has full rank for all $n \geq n_0$ for some $n_0 \in \mathbb{N}$.

We will restrict ourselves to natural link functions and refer to Section 4 of [FK85, p. 360] for the situation with non-natural link functions. For natural link functions, it holds that $u = \mathrm{id}$. When the observations in a generalized linear model are distributed with respect to some natural exponential family with density $p_\theta(y)$, the log-likelihood of an independent sample $Y_1 = y_1, \ldots, Y_n = y_n$ observed at $x_1, \ldots, x_n$ is

$$l_n(y_1, \ldots, y_n, \beta) = \sum_{i=1}^n (f(x_i)^T\beta\, y_i - b(f(x_i)^T\beta)) + c, \tag{2.3.1}$$

where $c = \sum_{i=1}^n \log(h(y_i))$ does not depend on β. The likelihood function gives rise to the maximum likelihood estimator [Sha03, Definition 4.3]:

Definition 2.11 (Maximum likelihood estimator)**.** Let $\hat{\beta}$ be a measurable function of some random variables $Y_1, \ldots, Y_n$. If the corresponding log-likelihood $l_n(y_1, \ldots, y_n, \beta)$ attains a local maximum over the parameter space B in $\hat{\beta}$, we say that $\hat{\beta}$ is a maximum likelihood estimator (MLE).

Definition 2.12 (Information matrix)**.** Let $Y_1, \ldots, Y_n$ be some random variables and $l_n(y_1, \ldots, y_n, \beta)$ the corresponding log-likelihood function. This yields the score function $s_n(y_1, \ldots, y_n, \beta)$ and the information matrix $M(x_1, \ldots, x_n, \beta)$:

$$\begin{aligned} s_n(y_1, \ldots, y_n, \beta) &:= \frac{\partial}{\partial\beta} l_n(y_1, \ldots, y_n, \beta), \\ M(x_1, \ldots, x_n, \beta) &:= \mathrm{Cov}_\beta(s_n(Y_1, \ldots, Y_n, \beta)). \end{aligned}$$

The information matrix of a vector of independent observations is the sum of information matrices of single observations [Sha03, Prop. 3.1]:

Proposition 2.13. *Let $Y = (Y_1, \ldots, Y_n)$ be a vector of independent random variables observed at $x_1, \ldots, x_n$. Then,*

$$M(x_1, \ldots, x_n, \beta) = \sum_{i=1}^n M(x_i, \beta), \tag{2.3.2}$$

where $M(x_i, \beta)$ is the information matrix of Y_i for $1 \leq i \leq n$.

We define $\mu_i(\beta) = \mu(f(x_i)^T\beta)$ and $\Sigma_i(\beta) = \Sigma(f(x_i)^T\beta)$. Then, we see that

$$s_n(y_1, \dots, y_n, \beta) = \sum_{i=1}^{n} f(x_i)(y_i - \mu_i(\beta)),$$
$$M(x_1, \dots, x_n, \beta) = \sum_{i=1}^{n} f(x_i)\Sigma_i(\beta)f(x_i)^T.$$

2.4. Cramér–Rao bound

In this section, we use the notation $M(\theta)$ for the information matrix of a random variable $Y = (Y_1, \dots, Y_n)$, as this section does not require Y to be distributed according to an exponential family. The parameter θ is chosen from a parameter space $\Theta \in \mathbb{R}^p$ such that the information matrix is a $p \times p$-matrix. The information matrix yields a strong implication for unbiased estimators in the form of the information inequality known as the Cramér–Rao bound. It states, that the covariance of an unbiased estimator $\hat{\Psi}(Y)$ for a function $\psi(\theta)$ of the parameter can be bounded below by the inverse of the information matrix [Sha03, Theorem 3.3]. Note that for matrices M_1, M_2, the notation $M_1 \geq M_2$ corresponds to the Loewner order, such that $M_1 \geq M_2$ means that $M_1 - M_2$ is positive semidefinite.

Theorem 2.14 (Cramér–Rao bound)**.** *Let $Y = (Y_1, \dots, Y_n)$ be distributed according to a probability measure W_θ where θ is chosen from the parameter space $\Theta \subset \mathbb{R}^p$. Let $\hat{\Psi}(Y)$ be an unbiased estimator for $\Psi(\theta)$ where $\Psi : \mathbb{R}^p \to \mathbb{R}^p$ is a differentiable function. Given a probability density $p_\theta = \frac{dW_\theta}{d\nu}$ with respect to a dominating measure ν for all $\theta \in \Theta$ that satisfies*

$$\frac{\partial}{\partial\theta}\int \kappa(y)p_\theta(y)\nu(dy) = \int \kappa(y)\frac{\partial}{\partial\theta}p_\theta(y)\nu(dy)$$

for $\kappa(y) = 1$ and $\kappa(y) = T(y)$ for all $\theta \in \Theta$ and assuming Θ° to be non-empty, it holds that

$$Cov(\hat{\Psi}(Y)) \geq \left(\frac{\partial}{\partial\theta}\Psi(\theta)\right)^T M(\theta)^{-1}\frac{\partial}{\partial\theta}\Psi(\theta).$$

If $\Psi(\theta) = \theta$, this simplifies to

$$Cov(\hat{\theta}) \geq M(\theta)^{-1}.$$

The Cramér–Rao bound was the first information inequality that was discovered and is therefore one of the origins of information theory and information geometry. [AJLS17, Theorem 5.7] displays a general Cramér–Rao bound in the terminology of information geometry that also holds for biased estimators, see also [AJLS17, Section 5.2.3]. For some special but important cases as for example linear models, the Cramér–Rao bound is realized as an equation for the least squares estimator (which is identical to the MLE for linear regression if we assume Gaussian errors). In these cases, the covariance of the MLE is therefore given as the inverse of the corresponding information matrix:

Example 2.15 (Linear regression realizes the Cramér–Rao bound)**.** Let $y_i = f(x)^T\theta+\varepsilon_i$ be a linear regression model with normally distributed errors $\varepsilon_i \sim \mathrm{N}(0,1)$, regression vector $f(x) = (1, x_1, \ldots, x_{p-1})$ and parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Then, the least-squares estimator $\hat{\theta}$ equals the MLE and realizes the Cramér-Rao bound as an equation:

$$\mathrm{Cov}(\hat{\theta}) = M(\theta)^{-1}.$$

Unfortunately, the equality above does not hold in general. Furthermore, for generalized linear models the MLE is not always unbiased, so Theorem 2.14 does not apply. Nevertheless, we will see in the following section that, under the right assumptions, the covariance of the MLE coincides asymptotically with the inverse information matrix.

2.5. Consistency and asymptotic normality of the MLE for GLM

This section gives a brief overview of the results on consistency and asymptotic normality of the MLE for GLM that were proved by Fahrmeir and Kaufmann in [FK85]. An estimator $\hat{\Psi}(Y)$ for a function Ψ of the parameter β_0 to be estimated from data $Y = (Y_1, \ldots, Y_n)$ in an experiment is *consistent* when it converges towards $\Psi(\beta_0)$ for $n \to \infty$. If the convergence is in probability this is denoted as *weak consistency* while almost sure convergence corresponds to *strong consistency*.

Theorem 2.16 ([FK85, Theorem 1])**.** *Under certain regularity assumptions, there is a sequence $(\hat{\beta}_n)_n$ of random variables with*

1. $\mathbb{P}(s_n(Y_1, \ldots, Y_n, \hat{\beta}) = 0) \to 1$,
2. $\hat{\beta} \xrightarrow{\mathbb{P}} \beta_0$.

The theorem shows the *weak consistency* of $\hat{\beta}_n$, so the convergence in probability of the estimator towards the true parameter. Furthermore, it is obtained that the score function converges in probability towards zero. With a slight modification of the regularity assumptions (see [FK85]), Fahrmeir and Kaufmann obtained almost sure convergence:

Theorem 2.17 ([FK85, Theorem 2])**.** *Under certain regularity assumptions, there is a sequence $(\hat{\beta}_n)_n$ of random variables and some random integer N_2 with*

1. $\mathbb{P}(s_n(Y_1, \ldots, Y_n, \hat{\beta}) = 0 \text{ for all } n \geq N_2) \to 1$,
2. $\hat{\beta} \xrightarrow{a.s.} \beta_0$.

The almost sure convergence of $\hat{\beta}$ towards β_0 is called *strong consistency*.
The following lemma provides the foundation for the result in Theorem 2.19.

Lemma 2.18 ([FK85, Lemma 1]). *Under certain regularity assumptions, the normed score function is asymptotically normal:*

$$M(x_1, \dots, x_n, \beta_0)^{-\frac{1}{2}} s_n(Y_1, \dots, Y_n, \beta_0) \xrightarrow{D} N(0, I).$$

Theorem 2.19 ([FK85, Theorem 3]). *Under the regularity assumptions of [FK85, Lemma 1], the normed MLE is asymptotically normal:*

$$M(x_1, \dots, x_n, \beta_0)^{\frac{1}{2}} (\hat{\beta} - \beta_0) \xrightarrow{D} N(0, I).$$

Additionally assuming the convergence of $\frac{1}{n} M(x_1, \dots, x_n, \beta)$ towards a limiting information matrix $\tilde{M}(\beta_0)$, we obtain

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \tilde{M}(\beta_0)^{-1}).$$

Therefore, the covariance of the MLE is asymptotically equal to the inverse of the information matrix. This allows us to make similar statements regarding the theory of design of experiments for GLM as for linear regression. For linear regression, an optimal design that "maximizes" the information matrix equivalently "minimizes" the covariance of the corresponding estimator. Now, with the results on consistency and asymptotic normality of the MLE for GLM, we can say that asymptotically, this is also true for GLM in the sense that the covariance of the asymptotic distribution is minimized. We are therefore able to introduce experimental designs for GLM in the following section.

2.6. Experimental design for generalized linear models

Section 2.4 established the connection between the information matrix and the covariance of an unbiased estimator. Section 2.5 shows that asymptotically, the inverse covariance of the MLE for GLM coincides with the information matrix. Now, it is customary to rate the MLE by the value of a risk function. Conveniently, the most common risk function is the covariance. In this setting, this means that the quality of the estimator in the sense of minimizing its covariance is related to finding an experiment with maximal information. Furthermore, studying the information matrix instead of the covariance has some advantages. For example, the information matrix of a vector of independent observations $Y = (Y_1, \dots, Y_n)$ is the sum of the corresponding information matrices of Y_i for $1 \leq i \leq n$ (cf. Proposition 2.13). This allows us to apply methods from convex optimization, as we will see later.

We assume a generalized linear model with observations $Y_1, \dots, Y_n$ distributed according to a one-parametric natural exponential family and a natural link function g, such that

$$\mathbb{E}_\beta(Y_i) = g^{-1}(f(x_i)^T \beta). \tag{2.6.1}$$

Now, $f(x_i)$ is a p-dimensional regression vector whose entries determine the mean of Y_i. The corresponding coordinates x_i which determine the entries of $f(x_i)$ are called *design*

points, while the set $\mathcal{X}$ of all possible design points is the *design space*. Let $\lambda_\beta(x)$ denote the *intensity function* such that $\Sigma(f(x)^T\beta) = \lambda_\beta(x)I_p$, where I_p is the $p \times p$ identity matrix. This holds as we assumed the corresponding natural exponential family to be one-parametric. Let $f_\beta(x) = \sqrt{\lambda_\beta(x)}f(x)$ denote the rescaled regression vector.

Definition 2.20 ((Induced) design space)**.** The set of all possible design points is the *design space* $\mathcal{X}$. In a generalized linear model with a regression function $f(x)$ and intensity $\lambda_\beta(x)$ for $x \in \mathcal{X}$, we call $\mathcal{U} = f_\beta(\mathcal{X})$ the *induced design space*.

The optimization procedure in optimal design is to assign proportions of observations to each design point, so that the resulting experiment maximizes the information matrix. In other words, we define a discrete probability measure ξ on $\mathcal{X}$, where $n\,\xi(x)$ is a non-negative integer. ξ is then called an *experimental design*, which is often abbreviated to *design*.

Definition 2.21 (Experimental design)**.** An *experimental design* of sample size n is a discrete probability measure ξ on some design space $\mathcal{X}$, so that $\xi(x) \in \frac{\mathbb{N}_0}{n}$ for all $x \in \mathcal{X}$. We call $\xi(x)$ the *design weight* at *design point* $x \in \mathcal{X}$.

This means that a design of sample size n can be described by a finite set of mutually distinct settings x_i, $i = 1, ..., m$, for the explanatory variable and the corresponding numbers $n_i \in \mathbb{N}$ of replications at x_i, where x_i is chosen from the design region $\mathcal{X}$ of potential settings. This means that we can write $\xi(x_i) = \frac{n_i}{n}$.

The information of a single design point $x \in \mathcal{X}$ is

$$M(x, \beta) = \lambda_\beta(x)f(x)f(x)^T.$$

We call $M(x, \beta)$ the *elemental information matrix* [AFHZ14] at design point x. With Proposition 2.13, we say that the information matrix of a design ξ is given as the sum of the elemental information matrices normalized by the number of observations, that is

$$M(\xi, \beta) = \sum_{x \in \mathcal{X}} \xi(x)M(x, \beta).$$

The information matrix of a design is a convex combination of elemental information matrices. This leads to the definition of the information matrix polytope:

Definition 2.22 (Information matrix polytope)**.** The convex hull

$$\mathcal{M}(\beta) = \text{conv}\{M(x, \beta) : x \in \mathcal{X}\}$$

is called the information matrix polytope.

We continue the example on logistic regression (cf. Example 2.10):

Example 2.23 (Logistic regression for binary responses, continued)**.** For the logistic regression, we obtain

$$\lambda_\beta(x) = \Sigma(f(x)^T\beta)$$

$$
\begin{aligned}
&= \frac{\partial^2}{\partial\theta^2} b(\theta)|_{\theta=f(x)^T\beta} \\
&= \frac{e^\theta}{(1+e^\theta)^2}|_{\theta=f(x)^T\beta} \\
&= \frac{e^{f(x)^T\beta}}{(1+e^{f(x)^T\beta})^2}
\end{aligned}
$$

Therefore, the information matrix of a design ξ is

$$M(\xi,\beta) = \sum_{x\in\mathcal{X}} \xi(x)\lambda_\beta(x) f(x) f(x)^T.$$

2.7. Approximate designs

Motivated from applications, we defined an experimental design to be a probability measure ξ, so that $n\,\xi(x)$ is a non-negative integer for all $x \in \mathcal{X}$. While this is convenient and necessary in applications, it has some analytical disadvantages for the theoretical approach, as the discontinuity is problematic. Therefore it is customary to drop the requirement for $n\,\xi(x)$ to be a non-negative integer and instead allow ξ to be any discrete probability measure with finite support on $\mathcal{X}$. This concept is known as approximate design [Kie74]:

Definition 2.24 (Approximate Design)**.** A discrete probability measure ξ with finite support on a design space $\mathcal{X}$ is an *approximate* experimental design.

Naturally, every point in the information matrix polytope corresponds to some design ξ and vice versa. The transition to approximate designs allows to apply methods from convex optimization to the problems in the theory of optimal design, as we replaced variables of the form $\frac{\mathbb{N}}{n}$ with continuous variables. This is the standard approach in design of experiments. In practice, an optimal design is derived from an optimal approximate design via careful rounding. Unfortunately, in some cases this might not provide the optimal design among the exact designs, that is those designs that satisfy $n\,\xi(x) \in \mathbb{N}$ for all $x \in \mathcal{X}$. If $\mathcal{X}$ is a finite set, the set Ξ containing all approximate designs is the subset of $\mathbb{R}^p$ known as the probability simplex:

Definition 2.25 (Probability Simplex)**.** The set of all designs on a given finite design space $\mathcal{X}$ is the probability simplex (or design simplex)

$$\Xi = \left\{ \xi \in \mathbb{R}^{|\mathcal{X}|}_{\geq 0} : \sum_{i=1}^{|\mathcal{X}|} \xi_i = 1 \right\}.$$

2.8. Optimality criteria

As we established before, the optimization step of experimental design is to maximize the information matrix. To do this, the experimenter chooses a criterion, which usually

is a map from the information matrix polytope to the real line, and searches for the design that maximizes the chosen criterion. This section, following [Sil80, Section 2.2], provides a brief introduction to two of the most common criteria with special attention to the D-criterion, which we will later use in this work.

Definition 2.26 (Optimality criterion)**.** An optimality criterion ϕ is a monotonic increasing functional $\phi : \mathcal{M} \to \mathbb{R}$ that maps from the information matrix polytope to the real line. The monotonicity is with respect to the Loewner order where $M_1 \geq M_2$ means that $M_1 - M_2$ is positive semidefinite. An optimality criterion then suffices $\phi(M_1) \geq \phi(M_2)$ for $M_1 \geq M_2$. If the information matrix depends on the parameter β, we write ϕ_β for the criterion function. We say that a criterion is *homogeneous* when for every $M \in \mathcal{M}$ and $a \in \mathbb{R}$ it holds that $\phi(aM) = a\phi(M)$.

The definition leads to the introduction of "quality measures" for designs, as one is interested in rating the optimality of a design.

Definition 2.27 (Efficiency)**.** Let ξ^* denote the optimal approximate design with respect to some homogeneous criterion ϕ. Then,

$$\text{eff}(\xi, \beta) = \frac{\phi_\beta(\xi)}{\phi_\beta(\xi^*)}$$

defines the efficiency of a design ξ [Sil80, p. 58].

It is customary to choose the homogenized versions of the functional $\phi_\beta(\xi)$ in the efficiency. This is done, as the resulting number is the inverse of the factor of the number of observations needed to achieve the same information as for the optimal design. For example, if the efficiency of some design is $\frac{1}{2}$, we need twice as many observations to obtain the same information as in the optimal case.

2.8.1. D-criterion

We introduce the D-criterion by choosing the logarithm of the determinant of the information matrix as the functional to maximize. The logarithm is taken to simplify computations.

Definition 2.28. An experimental design ξ^* is (locally) *D-optimal*, if

$$\log\det(M(\xi^*, \beta)) \geq \log\det(M(\xi, \beta))$$

for all $\xi \in \Xi$.

The functional $\log\det(M(\xi, \beta))$ is a suitable choice for an optimality criterion, as the determinant displays properties of the maximum likelihood estimator. For example, the volume of a confidence ellipsoid corresponds to the determinant of the covariance of the estimator. Due to the inverse relation between the covariance of the MLE and

the information matrix, maximizing the information coincides with minimizing the volume of confidence ellipsoid. The homogeneous D-criterion is $\phi_\beta(\xi) = (\det(M(\xi,\beta)))^{\frac{1}{p}}$. Therefore, the efficiency with respect to a D-optimal design ξ^* of some design ξ with $p = \dim(M(\xi^*))$, assuming that the regression functions are linearly independent on $\mathcal{X}$, is given as

$$\text{eff}_D(\xi,\beta) = \left(\frac{\det(M(\xi,\beta))}{\det(M(\xi^*,\beta))}\right)^{\frac{1}{p}}. \tag{2.8.1}$$

2.8.2. G-criterion

For a fixed $c \in \mathbb{R}^p$, it holds that the variance of the least-squares estimator of $c^T\beta$ associated to a design ξ is proportional to $c^T M(\xi,\beta)^{-1}c$. A criterion would therefore be established by choosing some $\mathcal{C} \subset \mathbb{R}^p$, so that a design ξ^* is optimal when it minimizes

$$\max_{c\in\mathcal{C}} c^T M(\xi,\beta)^{-1}c.$$

If $\mathcal{C} = f(\mathcal{X})$ is chosen as the induce design region such that

$$\xi^* = \arg\min_{\xi\in\Xi}\max_{x\in\mathcal{X}} f(x)^T M(\xi,\beta)^{-1} f(x),$$

we say that ξ^* is G-optimal. G-optimality is a standard example of a minimax criterion (see also [Fed72, p. 63]).

2.9. Local optimality and maximin designs

The optimality of an experimental design ξ is determined by the optimality criterion that is applied to the information matrix $M(\xi,\beta)$ of the design. In general, the optimality is hereby not only depending on the design itself but also on the parameter vector β. We therefore say that an experimental design ξ is *locally* optimal with respect to β and some optimality criterion. In special but important cases, the information matrix $M(\xi,\beta)$ does not depend on β, which means that an optimal design is globally optimal, i.e. optimal for all β. An example for this are linear models. The dependence on β has severe consequences on the process of finding an optimal design, as we need to choose a parameter to fix the optimal design. As the information varies throughout the β-space, the efficiency of a design is also dependent on β. This means that without prior knowledge of the true β, a practitioner is often interested in a design that is considered to be *maximin*-optimal. The term maximin-optimality, in different notions, is widely used in the theory of optimal design, so we fix the definition for this work:

Definition 2.29 (Maximin design)**.** A design ξ_r is a *maximin*-design, when it maximizes the minimal efficiency in a subset B' of the parameter space $B \subseteq \mathbb{R}^p$, so

$$\xi_r = \arg\max_{\xi\in\Xi}\min_{\beta\in B'} \text{eff}(\xi,\beta).$$

The choice of the parameter for the design can be done in various ways. The most common are the following: Make a pilot study according to a maximin-design (in many models for example a uniform design) to find a rough estimate for β. Alternatively, it is common to use expert knowledge to select β. This approach can also be implemented as a sequential method, that adapts the design based on the gained information. Unfortunately, this means that the observations are not independent anymore, such that the definition for the information matrix changes. These problems has been researched since at least 1953 [Che53] and is an ongoing research topic. We refer the reader to [DMP+04] for further details and references.

2.10. Equivalence theorems

Section 2.8 introduced the various optimality criteria, which may be applied to find optimal experimental designs. This section displays an equivalence theorem that connects D-optimality to convex optimization. The description as equivalence theorem stems from the fact that it shows the equivalence of D-optimality and G-optimality (see Section 2.8.2 and Eq. (2.10.1)). For a general introduction, see [Puk93, Sil80]. Now, let $\phi(M) = \log\det(M)$ denote the D-criterion. The following definition is taken from [Sil80, Section 3.5.2].

Definition 2.30 (Directional derivative)**.** The *directional derivative* (*Fréchet derivative*) of the D-optimality criterion at M_1 in the direction of M_2 for some $(m-1)\times(m-1)$-matrices M_1, M_2 is

$$F_D(M_1, M_2) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \left(\log\det((1-\varepsilon)M_1 + \varepsilon M_2) - \log\det(M_1)\right).$$

It holds for $M_1, M_2 \in \mathcal{M}$ that $(1-\varepsilon)M_1 + \varepsilon M_2 \in \mathcal{M}$, as the information matrix polytope $\mathcal{M}$ is convex. Therefore, $\log\det((1-\varepsilon)M_1 + \varepsilon M_2)$ is defined. Through the concavity of $\log\det$ it follows that

$$\frac{1}{\varepsilon} \left(\log\det((1-\varepsilon)M_1 + \varepsilon M_2) - \log\det(M_1)\right)$$

is a non-increasing function of ε in $0 < \varepsilon \leq 1$ and therefore $F_D(M_1, M_2)$ always exists if we allow for $F_D(M_1, M_2) = +\infty$ [Sil80, Section 3.5.2 (ii)].

As the D-criterion is a concave function on a convex set (the information matrix polytope), its maximum can be interpreted as the "top of the hill", which means that a design is D-optimal, when its directional derivative to all matrices in the information matrix polytope is non-positive. Now, if the hill is smooth, the problem may be reduced to only the directional derivatives towards the elemental information matrices [Sil80, Theorem 3.7]:

Theorem 2.31 (Kiefer–Wolfowitz equivalence theorem)**.** *ξ^* is D-optimal if and only if $F_D(M(\xi^*), f(x)f(x)^T) \leq 0$ for all $x \in \mathcal{X}$.*

The following corollary is a consequence of the Kiefer–Wolfowitz equivalence theorem:

Corollary 2.32 ([Sil80, Corollary 3.10]). *For a D-optimal design ξ^* it holds that $F_D(M(\xi^*), f(x)f(x)^T) = 0$ for all $x \in \mathcal{X}$ with $\xi^*(x) > 0$.*

This means that for a D-optimal design the directional derivatives towards the elemental information matrices is zero for support points of the design. Silvey derives the following simplification for the directional derivative for the D-criterion:

$$\begin{aligned}\log\det((1-\varepsilon)M_1 + \varepsilon M_2) - \log\det(M_1) &= \log\det((1-\varepsilon)I + \varepsilon M_2 M_1^{-1}) \\ &= \log(1 + \varepsilon\ \mathrm{tr}(M_2 M_1^{-1} - I)) + O(\varepsilon^2) \\ &= \varepsilon\,\mathrm{tr}(M_2 M_1^{-1} - I) + O(\varepsilon^2).\end{aligned}$$

Here, $\mathrm{tr}(M)$ denotes the trace of some matrix M. These considerations imply that

$$\begin{aligned}F_D(M_1, M_2) &= \mathrm{tr}(M_2 M_1^{-1} - I) \\ &= \mathrm{tr}(M_2 M_1^{-1}) - \dim(M_1).\end{aligned}$$

For the directional derivatives in Theorem 2.31 it follows that

$$F_D(M(\xi^*), f(x)f(x)^T) = f(x)^T M(\xi^*)^{-1} f(x) - p. \tag{2.10.1}$$

Theorem 2.31 is only formulated for linear models. To extend this to GLM as introduced in Section 2.6, we transfer the information matrix

$$M(\xi, \beta) = \sum_{x \in \mathcal{X}} \xi(x)\lambda_\beta(x) f(x) f(x)^T$$

into a pointwise linear model with the rescaled regression function $f_\beta(x) = \sqrt{\lambda_\beta(x)} f(x)$. We write

$$M_\beta(\xi) = M(\xi, \beta) = \sum_{x \in \mathcal{X}} \xi(x) f_\beta(x) f_\beta(x)^T$$

and obtain the extended Kiefer–Wolfowitz equivalence theorem:

Theorem 2.33 (Extended Kiefer–Wolfowitz equivalence theorem, see [Fed72, Theorem 2.2.1]). *ξ^* is locally D-optimal in β if and only if $F_D(M_\beta(\xi^*), f_\beta(x) f_\beta(x)^T) \le 0$ for all $x \in \mathcal{X}$.*

The following corollary is a consequence of the extended Kiefer–Wolfowitz equivalence theorem:

Corollary 2.34 ([Sil80, Corollary 3.10]). *It holds that $F_D(M_\beta(\xi^*), f_\beta(x) f_\beta(x)^T) = 0$ for all $x \in \mathcal{X}$ with $\xi(x) > 0$.*

As a consequence of Eq. (2.10.1), it follows that

$$\begin{aligned}F_D(M_\beta(\xi^*), f_\beta(x) f_\beta(x)^T) &= F_D(M(\xi^*, \beta), \lambda_\beta(x) f(x) f(x)^T) \\ &= \lambda_\beta(x) f(x)^T M(\xi^*, \beta)^{-1} f(x) - p.\end{aligned} \tag{2.10.2}$$

Therefore, the result of Theorem 2.33 is that a design ξ^* is locally D-optimal in β if and only if $\lambda_\beta(x) f(x)^T M(\xi^*, \beta)^{-1} f(x) \le p$ for all $x \in \mathcal{X}$.

2.11. Semi-algebraic geometry of D-optimal experimental designs for GLM

To explain the algebraic perspective on optimal design, we begin by introducing some basic definitions from algebraic geometry [CLO15, BCR13]. Let $R = \mathbb{K}[x_1, \ldots, x_n]$ be a polynomial ring in the variables $x_1, \ldots, x_n$ over some field $\mathbb{K}$. A monomial is a product of the form

$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$$

with exponents $\alpha_1 \ldots \alpha_n \in \mathbb{N}_0$. A polynomial f is a finite linear combination of monomials with coefficients $a_\alpha \in \mathbb{K}$, so

$$f = \sum_\alpha a_\alpha x^\alpha.$$

The zero set of a collection of polynomials is an *affine variety*:

Definition 2.35 (Affine variety)**.** Let $f_1, \ldots, f_s$ be polynomials in R. The set

$$\mathrm{V}(f_1, \ldots, f_s) = \{a \in \mathbb{K}^n : f_i(a) = 0 \text{ for all } 1 \leq i \leq s\}$$

is the affine variety of $f_1, \ldots, f_s$.

In experimental design, the equivalence theorem often translates the optimization problem into a set of polynomial inequalities and equations. As varieties only include polynomial equations, the addition of inequalities needs a new definition. A subset of $\mathbb{K}^n$ defined by polynomial inequalities and equations is a *semi-algebraic set*:

Definition 2.36 (Semi-algebraic set, see [BCR13, Def. 2.1.4])**.** Let f_{ij} be polynomials in R and $*_{ij}$ be either $<$ or $=$ for $1 \leq i \leq s$ and $1 \leq j \leq r_i$ with $s, r_i \in \mathbb{N}$. A subset of $\mathbb{K}^n$ of the form

$$\bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{K} : f_{ij} \ *_{ij} \ 0\}$$

is a *semi-algebraic set*.

A useful property of semi-algebraic sets is that there always exists a certificate to determine if they are empty by the Positivstellensatz from real algebraic geometry (see [BCR13]). This means that if the inequality system has no solution, then one can combine the inequalities to produce an explicit contradiction. There are computational tools to search for such certificates. These separate into numerical and exact approaches. An example for a numerical tool is SOStools [PAV+13], which approximates the correct contradiction via semidefinite programming. A symbolic approach is quantifier elimination (QE), which is a symbolic algorithm that finds a quantifier-free description of any semi-algebraic set. QE is a direct consequence of the Tarski–Seidenberg theorem, see [BCR13, Section 5.2]. Unfortunately, QE is of doubly-exponential computation time, therefore

the problems where we can apply QE cannot be "too big". QE is for example implemented in the `Reduce` functionality of MATHEMATICA. Examples for the application QE can be found in Chapter 3.

The Kiefer–Wolfowitz equivalence theorem shows that a design ξ^* is D-optimal if and only if all directional derivatives of its information matrix towards the elemental information matrices are negative. In (2.10.2), the directional derivatives towards the elemental information matrices are given as

$$\lambda_\beta(x) f(x)^T M(\xi^*, \beta)^{-1} f(x)^T - p \leq 0,$$

where $p = \dim(M(\xi^*, \beta))$. If a matrix is invertible, its inverse can be computed via the adjugate matrix:

Definition 2.37 (Adjugate matrix)**.** Let M be a $m \times m$-matrix. The adjugate $\operatorname{adj}(M)$ of M is the transpose of the cofactor matrix

$$\operatorname{adj}(M)^T = \left((-1)^{i+j} M^{ij}\right)_{1 \leq i,j \leq m},$$

where M^{ij} denotes the (i,j)-minor of M, so the determinant of the submatrix of M generated by eliminating the i-th row and j-th column. By Cramér's rule, it follows that

$$M^{-1} = \frac{\operatorname{adj}(M)}{\det(M)}.$$

For generalized linear models, under the assumption that the entries of $f(x)$ are polynomials, the term $f(x_i)^T M(\xi^*, \beta)^{-1} f(x_i)^T - p$ is a rational function in the variables $x_i, w_i := \xi(x_i)$ and $\lambda_i := \lambda_\beta(x_i)$, as it holds that

$$M(\xi^*, \beta)^{-1} = \frac{\operatorname{adj}(M(\xi^*, \beta))}{\det(M(\xi^*, \beta))},$$

where the adjugate matrix entries consist only of polynomials in said variables. Assuming that $\det(M(\xi^*, \beta)) \neq 0$, as otherwise ξ^* could not be D-optimal, this implies that the region in the parameter space where ξ^* is D-optimal is a semi-algebraic set.

For a discrete design space $\mathcal{X}$, a design ξ^* is therefore D-optimal if and only if

$$\lambda_\beta(x) f(x)^T \operatorname{adj}(M(\xi^*, \beta)) f(x)^T - p \det(M(\xi^*, \beta)) \leq 0 \tag{2.11.1}$$

for all $x \in \mathcal{X}$. If $\xi^*(x) > 0$, the inequality in (2.11.1) is realized as an equation (cf. Corollary 2.34). This implies that the optimality region of a design ξ^* with $\xi^*(x) > 0$ for all $x \in \mathcal{X}$ is an affine variety intersected with the constraints imposed by the model on the parameters and by ξ^* being a design.

3. The semi-algebraic geometry of the Bradley–Terry model

This chapter discusses optimal designs and the geometry of their optimality regions in the parameter space for the Bradley–Terry paired comparison model. Section 3.1 introduces the model. Our main result is presented in Section 3.2. Furthermore, we exhibit the geometry of designs with minimal support for an arbitrary number of alternatives in Section 3.3 and study the special case for four alternatives in Section 3.4. The chapter ends with a discussion in Section 3.5.

3.1. General setup

We consider pairs (i, j) of alternatives $i, j = 1, \ldots, m$. The preference of i over j is modeled by a binary variable $Y(i, j)$ taking the value $Y(i, j) = 1$ if i is preferred over j and $Y(i, j) = 0$ otherwise. We do not consider any order effects here. The main assumption of the Bradley–Terry model is that there is a hidden ranking of the alternatives according to some numerical preference value $\pi_i > 0$, $i = 1, \ldots, m$. When presented with the pair (i, j), the probability of preferring i over j is

$$\mathbb{P}(Y(i, j) = 1) = \frac{\pi_i}{\pi_i + \pi_j}.$$

The model can be transformed into a logistic regression model using $\beta_i := \log(\pi_i)$. Then

$$\mathbb{P}(Y(i, j) = 1) = \frac{1}{1 + \exp(-(\beta_i - \beta_j))} = g^{-1}(\beta_i - \beta_j)$$

with $g^{-1}(z) = (1 + \exp(-z))^{-1}$ as the inverse logit link function (cf. Example 2.10).

Scaling all π_i with a constant factor leaves the preference probabilities invariant. Therefore one can without loss of generality assume that $\pi_m = 1$ or $\beta_m = 0$. This means that the number of parameters of the Bradley–Terry model is $m - 1$. The number of alternatives minus 1 is the main measure of complexity of the design theory for the Bradley–Terry model as it equals the dimension of the induced design space. The remaining parameters can be identified and $\beta_m = 0$ is known as *control coding*. We denote by e_i the i-th standard unit vector in $\mathbb{R}^{m-1}$. To exhibit our model as a generalized linear model, the *regression vector* for a pair (i, j) is

$$f(i, j) = \begin{cases} e_i - e_j, & \text{for } i, j \neq m, \\ e_i, & \text{for } i < j, j = m, \\ 0 & \text{for } i = j = m. \end{cases}$$

With $\beta^T = (\beta_1, \dots, \beta_{m-1})$ this yields $\mathbb{P}(Y(i,j) = 1) = g^{-1}(f(i,j)^T\beta)$ where $f(i,j)^T\beta$ is the linear predictor. The design space of the Bradley–Terry paired comparison model is

$$\mathcal{X} = \{(i,j) \,:\, i,j = 1,...,m,\, i < j\}.$$

It consists of all pairs of ordered alternatives. The pairs (i,j) and (j,i) bear the same information, and the comparison (i,i) of two identical alternatives does not have any information at all (as can be seen easily later). Therefore, whenever there are two alternatives $i,j \in \{1,\dots,m\}$ we assume $i < j$. An experimental design ξ as introduced in Section 2.6 is an assignment of a weight $w_{ij} := \xi(i,j) \geq 0$ to each point $(i,j) \in \mathcal{X}$, such that $\sum_{ij} w_{ij} = 1$. We assume the designs to be approximate (see Section 2.7). The information gained from one observation of $Y(i,j)$ is encoded in the information matrix

$$M((i,j),\beta) = \lambda_{ij} f(i,j) f(i,j)^T \in \mathbb{R}^{(m-1)\times(m-1)},$$

where $\lambda_{ij} := \lambda_\beta(i,j) = \frac{e^{\beta_i-\beta_j}}{(1+e^{\beta_i-\beta_j})^2}$ is referred to as the *intensity* in [GS08], see also Example 2.23. It holds that $M((i,j),\beta) = M((j,i),\beta)$ and $M((i,i),\beta) = 0$. Assuming independent observations, the information matrix for a design ξ with weights w_{ij} is the $(m-1)\times(m-1)$-matrix

$$M(\xi,\beta) = \sum_{(i,j)} w_{ij} M((i,j),\beta) = \sum_{(i,j)} w_{ij}\lambda_{ij} f(i,j) f(i,j)^T. \tag{3.1.1}$$

As explained in Section 2.9, one speaks of local optimality, if the optimal choice of a design depends on the unknown parameters that one wants to learn about, see also [Che53]. The methods of convex optimization suggest to study the directional derivatives of the target function, cf. Definition 2.30. It is shown in [Sil80, Sections 3.8 and 3.11] (see Eq. (2.10.2)) that

$$F_D(M(\xi,\beta), M((i,j),\beta)) = \lambda_{ij} f(i,j)^T M(\xi,\beta)^{-1} f(i,j) - (m-1). \tag{3.1.2}$$

Theorem 2.33 yields that a design ξ^* is locally D-optimal if and only if

$$\lambda_{ij} f(i,j)^T M(\xi^*,\beta)^{-1} f(i,j) \leq m-1 \tag{3.1.3}$$

for all $1 \leq i < j \leq m$. Corollary 2.34 states that for design points (i,j) with positive weight in ξ^*, the inequalities (3.1.3) in Theorem 2.33 hold with equality. One of our main observations about the Bradley–Terry model is that it is useful to represent pairs (i,j) with positive weights w_{ij} as the edges of an undirected graph on the vertex set $\{1,\dots,m\}$.

Definition 3.1. A *graph representation of a design* ξ for the Bradley–Terry model is the undirected simple graph with vertex set $\{1,\dots,m\}$, and edge set $E = \{(i,j) : w_{ij} > 0\}$.

Using standard notions from graph theory, a *tree* is a connected graph with no cycles. A *path* is a tree in which every vertex is connected to at most two other vertices.

We exploit the symmetry of the model. The symmetric group $\mathrm{Sym}(m)$ of all bijective self-maps of $\{1,\dots,m\}$ permutes the alternatives. The permutation action extends to ordered pairs by acting on both entries of the pair simultaneously (and changing the order if necessary). The action also extends naturally to designs ξ on pairs (i,j) by putting, for any $\sigma \in \mathrm{Sym}(m)$, $(\xi^\sigma)_{(i,j)} = \xi_{\sigma^{-1}(i,j)}$. A graph representation of an entire orbit under this action is simply the unlabeled graph. Proposition 3.2 below expresses that for properties of the model it is irrelevant which alternative is alternative 1, which is alternative 2 and so on. One only needs to take care that upon relabeling the parameters, regression vectors, etc. are relabeled accordingly.

In our setup we have singled out the last alternative m and set $\beta_m = 0$ to have identifiable parameters. This changes the symmetry and needs to be accounted for. The concepts of this chapter, however, are compatible with this. For example the value of the determinant of a design is invariant:

Proposition 3.2. *Let $\sigma \in \mathrm{Sym}(m)$ and let ξ be any design. Then ξ is D-optimal for the parameters $\beta = (\beta_1,\dots,\beta_{m-1},0)$ if and only if ξ^σ is D-optimal for $Q_\sigma^{-T}\beta$, where $\sigma \mapsto Q_\sigma$ is a group homomorphism from $\mathrm{Sym}(m)$ to the group of invertible $(m-1)\times(m-1)$-matrices satisfying $f(\sigma(i),\sigma(j)) = Q_\sigma f(i,j)$ for all $\sigma \in \mathrm{Sym}(m)$.*

Proof. By [RS16, Section 2], the design ξ^σ is locally optimal for the parameter $Q_\sigma^{-T}\beta$ if and only if there exist matrices Q_σ as in the statement. As transpositions generate all permutations, it suffices to show the existence of such a Q_σ for all transpositions. For transpositions of $i < m$ and $j < m$, let Q_σ be the usual permutation matrix. For a transposition (im), let Q_σ equal an identity matrix, with the i-th row replaced by the row $(-1\dots,-1)$. Then, for an arbitrary permutation σ, it holds that $f(\sigma(i),\sigma(j)) = Q_\sigma f(i,j)$. □

3.2. Saturated designs and graph-representation

An experimental design is *saturated* if its support has cardinality equal to the number of free parameters of the model. In our case of D-optimality, if a design has support size strictly smaller than $m-1$, then the determinant of the information matrix vanishes and optimality is impossible. A useful result about saturated designs is that their weights are completely rigid: they are all equal (see [Sil80, Lemma 5.1.3]). We first study which saturated designs can be D-optimal. A saturated design has a quite restrictive structure on the observations, now expressed in terms of its graph representation. The following simple fact is reminiscent of the *connectedness* of block designs with block length two in [SS89, p.2].

Lemma 3.3. *For any D-optimal saturated design ξ of the Bradley–Terry paired comparison model, the graph representation of the support is a tree.*

Proof. A saturated design consists of $m-1$ equally weighted comparisons. If there is a cycle $i_1,\dots,i_k$ in the graph representation of the design, then there is at least one alternative that does not appear in the design and therefore represented by a disconnected

vertex in the graph representation. Now, the $(m-1)\times(m-1)$-information matrix of a saturated design is a sum of $m-1$ rank one matrices of the form $\lambda_{ij}f(i,j)f(i,j)^T$. For $1 \leq i < j \leq m-1$, these rank one matrices only have entries in the i-th and j-th row and column. For $j = m$, there is only one entry λ_{im} in the intersection of the i-th row and i-th column. Thereby, if a saturated design contains a cycle and thus misses one alternative, the information matrix has no non-zero entries in either the corresponding row or the corresponding column. Therefore the determinant of the information matrix is zero, and the design can never be optimal. By Proposition 3.2, this holds for all saturated designs that contain cycles. □

Based on this fact we can determine the saturated optimal designs for the Bradley–Terry model.

Theorem 3.4. *In the Bradley–Terry paired comparison model with m alternatives, every saturated optimal design's graph representation is a path on $[m] := \{1, 2, \dots, m\}$.*

Proof. Let ξ be a saturated design for the Bradley–Terry model with m alternatives. Without loss of generality, we assume that ξ has exactly one comparison that contains m. This holds, as an optimal saturated design is a tree. Let F be the (square) matrix of the transposed regression vectors of the design points

$$F = \begin{pmatrix} f(i_1, j_1)^T \\ f(i_2, j_2)^T \\ \vdots \\ f(i_{m-2}, j_{m-2})^T \\ f(i_{m-1}, m)^T \end{pmatrix},$$

and define $Q = \operatorname{diag}(\lambda_{i_1,j_1}, \dots, \lambda_{i_{m-1},m})$ as a diagonal matrix of intensities and correspondingly $W = \operatorname{diag}(w_{i_1,j_1}, \dots, w_{i_{m-1},m})$ for the weights of the design points. Then, the information matrix is $M(\xi, \beta) = F^T W Q F$, and inserting this into (3.1.2), we obtain the directional derivatives for every $1 \leq i < j \leq m-1$ as

$$\lambda_{ij} f(i,j)^T F^{-1} Q^{-1} W^{-1} F^{-T} f(i,j) - (m-1).$$

If the design is D-optimal, this formula is non-positive for every $1 \leq i < j \leq m-1$. Since all weights are equal to $\frac{1}{m-1}$ this is equivalent to

$$\lambda_{ij} f(i,j)^T F^{-1} Q^{-1} F^{-T} f(i,j) \leq 1.$$

The proof is by downward induction. To this end, we remove one alternative and its associated design point and show that the reduced design $\bar{\xi}$ is optimal on the reduced design space. Without loss of generality we can assume that the optimal design has only one comparison $(1, v)$ in which alternative 1 is involved. We can also assume that $v = 2$ using the $\mathrm{Sym}(m)$ symmetry and Proposition 3.2. We remove alternative 1. Consider the Bradley–Terry model on the alternatives $\{2, \dots, m\}$. Its information matrix is a product $\bar{F}\bar{W}\bar{Q}\bar{F}^T$, where $\bar{W}$ and $\bar{F}$ are the lower-right $(m-2)\times(m-2)$-submatrices

of $\frac{m-1}{m-2}W$ and F, respectively, and $\bar{Q}$ is the diagonal matrix of the reduced model's intensities $\bar{\lambda}_{ij}$. Through our assumptions,

$$F = \begin{pmatrix} 1 & -1 & 0 \cdots \cdots 0 \\ 0 & & \\ \vdots & & \bar{F} \\ 0 & & \end{pmatrix}.$$

We show the implication

$$\begin{aligned} &\lambda_{ij} f(i,j)^T F^{-1} Q^{-1} F^{-T} f(i,j) \leq 1 \quad \text{for all } 2 \leq i < j \leq m \\ \Rightarrow\ &\bar{\lambda}_{ij} \bar{f}(i,j)^T \bar{F}^{-1} \bar{Q}^{-1} \bar{F}^{-T} \bar{f}(i,j) \leq 1 \quad \text{for all } 2 \leq i < j \leq m. \end{aligned}$$

This implies that the design $\bar{\xi}$ with equal weights $\frac{1}{m-2}$ on $E \setminus \{1,2\}$ is optimal for the reduced model. Since $\bar{\lambda}_{ij} = \lambda_{ij}$, we only have to show

$$\bar{f}(i,j)^T \bar{F}^{-1} \bar{Q}^{-1} \bar{F}^{-T} \bar{f}(i,j) \leq f(i,j)^T F^{-1} Q^{-1} F^{-T} f(i,j) \tag{3.2.1}$$

for all $2 \leq i < j \leq m$. Now let

$$F^{-1} = \begin{pmatrix} a_{11} & a_{12}^T \\ a_{21} & A_1 \end{pmatrix}$$

for some $(m-2) \times (m-2)$-matrix A_1. This leads to $\bar{F}^{-1} = A_1 - \frac{1}{a_{11}} a_{21} a_{12}^T$. One checks $a_{21} = 0$, so that

$$F^{-1} = \begin{pmatrix} 1 & a_{12}^T \\ 0 & A_1 \end{pmatrix}.$$

This means, that $\bar{F}^{-1} = A_1$. Now, as $f(i,j)^T = (0, \bar{f}(i,j)^T)$,

$$\begin{aligned} f(i,j)^T F^{-1} Q^{-1} F^{-T} f(i,j) &= \begin{pmatrix} 0 & \bar{f}(i,j)^T \end{pmatrix} \begin{pmatrix} 1 & a_{12}^T \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} \frac{1}{\lambda_{12}} & \\ & \bar{Q}^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ a_{12} & A_1^T \end{pmatrix} \begin{pmatrix} 0 \\ \bar{f}(i,j) \end{pmatrix} \\ &= \begin{pmatrix} 0 & \bar{f}(i,j)^T A_1 \end{pmatrix} \begin{pmatrix} \frac{1}{\lambda_{12}} & \\ & \bar{Q}^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ A_1^T \bar{f}(i,j) \end{pmatrix} \\ &= \bar{f}(i,j)^T A_1 \bar{Q}^{-1} A_1^T \bar{f}(i,j) \\ &= \bar{f}(i,j)^T \bar{F}^{-1} \bar{Q}^{-1} \bar{F}^{-T} \bar{f}(i,j). \end{aligned}$$

In fact, (3.2.1) is realized as an equation and the reduced saturated design is optimal. Now, if ξ was not a path, iterating this procedure eventually leads to an optimal saturated design for Bradley–Terry model on four alternatives that is also not a path. Such a design does not exist by the explicit computations in Section 3.4. Hence, the graph representation of an optimal saturated design is a path. □

3.3. Optimality regions of saturated designs

We now describe the sets of parameters for which a given saturated design from Theorem 3.4 is optimal. We call such a set the *region of optimality* of the design. Knowing these regions simplifies the experimental design problem since it can be combined with prior knowledge about the parameters (e.g. from a pilot study). Also, knowing if the regions are big or small yields information about the maximin-optimality of designs.

Exploiting the symmetry in Proposition 3.2, it suffices to study a single design representing all saturated designs. This is the path $(1,2),(2,3),\dots,(m-1,m)$.

Lemma 3.5. *The optimality region of the design* $(12,23,34,\dots,(m-1)m)$ *is defined by the inequalities*

$$g(i,j)=\lambda_{ij}\sum_{k=i}^{j-1}\frac{1}{\lambda_{k(k+1)}}\leq 1,\qquad 1\leq i<m,\ i<j\leq m.$$

Furthermore, this region is not empty.

Proof. We apply Theorem 2.33 to find the optimality regions of the saturated design $(12,23,34,\dots,(m-1)m)$. Therefore, one has to analyze the directional derivatives $f(i,j)^TF^{-1}Q^{-1}F^{-T}f(i,j)-(m-1)$, where $f(i,j)$ are the regression vectors, Q is a diagonal matrix of the design intensities λ_{12},λ_{23}, …, $\lambda_{(m-1)m}$ and F is the matrix of the transposed regression vectors. So,

$$F=\begin{pmatrix}1&-1&&&\\&\ddots&\ddots&&\\&&\ddots&\ddots&\\&&&1&-1\\&&&&1\end{pmatrix}\qquad\text{and}\qquad F^{-1}=\begin{pmatrix}1&\cdots&\cdots&1\\&\ddots&&\vdots\\&&\ddots&\vdots\\&&&1\end{pmatrix}.$$

For $i<j<m$, $f(i,j)=e_i-e_j$. This leads to

$$f(i,j)^TF^{-1}=(\mathbb{1}_{\{i=1\}},\mathbb{1}_{\{i\leq 2<j\}},\mathbb{1}_{\{i\leq 3<j\}},\dots,\mathbb{1}_{\{i\leq m-2<j\}},0).$$

For $i<j=m$, $f(i,m)=e_i$. Therefore

$$f(i,m)^TF^{-1}=(\mathbb{1}_{\{i=1\}},\mathbb{1}_{\{i\leq 2\}},\mathbb{1}_{\{i\leq 3\}},\dots,\mathbb{1}_{\{i\leq m-2\}},1).$$

This means that the directional derivative in the direction (i,j) for $j<m$ is

$$\lambda_{ij}(m-1)(\mathbb{1}_{\{i\leq 1\}},\mathbb{1}_{\{i\leq 2<j\}},\dots,\mathbb{1}_{\{i\leq m-2<j\}},0)\begin{pmatrix}\frac{1}{\lambda_{12}}&&&\\&\frac{1}{\lambda_{23}}&&\\&&\ddots&\\&&&\frac{1}{\lambda_{(m-1)m}}\end{pmatrix}\begin{pmatrix}\mathbb{1}_{\{i\leq 1\}}\\\mathbb{1}_{\{i\leq 2<j\}}\\\vdots\\\mathbb{1}_{\{i\leq m-2<j\}}\\0\end{pmatrix}$$

$$=\lambda_{ij}(m-1)\sum_{k=1}^{m-2}\frac{\mathbb{1}_{\{i\leq k<j\}}}{\lambda_{k(k+1)}}$$

$$=\lambda_{ij}(m-1)\sum_{k=i}^{j-1}\frac{1}{\lambda_{k(k+1)}}$$

and for $j = m$

$$\lambda_{im}(m-1)(\mathbb{1}_{\{i\leq 1\}}, \mathbb{1}_{\{i\leq 2\}}, \dots, \mathbb{1}_{\{i\leq m-2\}}, 1) \begin{pmatrix} \frac{1}{\lambda_{12}} & & & \\ & \frac{1}{\lambda_{23}} & \ddots & \\ & & \ddots & \frac{1}{\lambda_{(m-1)m}} \end{pmatrix} \begin{pmatrix} \mathbb{1}_{\{i\leq 1\}} \\ \mathbb{1}_{\{i\leq 2\}} \\ \vdots \\ \mathbb{1}_{\{i\leq m-2\}} \\ 1 \end{pmatrix}$$

$$= \lambda_{im}(m-1) \sum_{k=1}^{m-1} \frac{\mathbb{1}_{\{i\leq k\}}}{\lambda_{k(k+1)}}$$

$$= \lambda_{im}(m-1) \sum_{k=i}^{m-1} \frac{1}{\lambda_{k(k+1)}}.$$

For $j = i+1$ the directional derivatives are 0 by (3.1.2) and Corollary 2.34. Let

$$g(i,j) = \lambda_{ij} \sum_{k=i}^{j-1} \frac{1}{\lambda_{k(k+1)}}.$$

By Theorem 2.33 the optimality region of the design $(12, 23, 34, \dots, (m-1)m)$ is given by $\{g(i,j) \leq 1 \,:\, 1 \leq i < j \leq m\}$. To exhibit a point in the optimality region, let $\beta_i = i\beta_1$ and thus $\pi_i = \pi_1^i$. This implies

$$\lambda_{ij} = \frac{\pi_1^{j-i}}{(1+\pi_1^{j-i})^2},$$

and therefore

$$g(i,j) = \frac{\pi_1^{j-i}}{(1+\pi_1^{j-i})^2} \sum_{k=i}^{j-1} \frac{(1+\pi_1)^2}{\pi_1} = \frac{(j-i)\pi_1^{j-i-1}(1+\pi_1)^2}{(1+\pi_1^{j-i})^2},$$

which is at most 1 for all $1 \leq i < j \leq m$ if just π_1 is sufficiently large. □

Theorem 3.6. *The optimality regions of all saturated designs corresponding to paths, i.e. of all optimal saturated designs, are in the* $\mathrm{Sym}(m)$*-orbit of the saturated design for* $(12, 23, 34, \dots, (m-1)m)$*. The optimality regions are defined by the inequalities*

$$\{g(\sigma(i), \sigma(j)) \leq 1 \,:\, 1 \leq i < m,\, i < j \leq m\}.$$

where $\sigma \in \mathrm{Sym}(m)$ *is a permutation turning* $(12, 23, 34, \dots, (m-1)m)$ *into the given path.*

Proof. Theorem 3.4 shows that the saturated optimal designs correspond to paths. By Proposition 3.2, we can choose any representative for the orbit of path designs. We choose $(12, 23, 34, \dots, (m-1)m)$ and plug in the results from Lemma 3.5. □

3.4. Explicit solutions for four alternatives

This section studies the optimal designs for the Bradley–Terry paired comparison model with four alternatives. We first deal with the case of saturated designs, i.e. optimal designs whose supports consist of only 3 design points. The unsaturated case with 4, 5 or 6 support points follows in Section 3.4.2.

The Bradley–Terry paired comparison model with 4 alternatives has 3 identifiable parameters $\beta_1, \beta_2, \beta_3$. As above we use $\beta_i := \log(\pi_i)$ and $\beta_4 = 0$. Our goal is to cover all of $\mathbb{R}^3$ with regions of optimality of specific explicit designs. The regression vectors for four alternatives are

$$\begin{aligned} f(1,2) &= (1,-1,0)^T, & f(1,3) &= (1,0,-1)^T, & f(1,4) &= (1,0,0)^T, \\ f(2,3) &= (0,1,-1)^T & f(2,4) &= (0,1,0)^T, & f(3,4) &= (0,0,1)^T. \end{aligned}$$

3.4.1. Saturated Designs

For saturated designs with non-singular information matrix, the optimality criterion in Theorem 2.33 yields a system of inequalities in the intensities λ_{ij}. We find these first. According to [Sil80, Lemma 5.1.3], a saturated design has three positive weights whose values are all $\frac{1}{3}$, the remaining weights being zero. There are $\binom{6}{3} = 20$ possible saturated designs. Exactly 16 of them have a non-singular information matrix. Among the 16, only 12 have a non-empty region of optimality. We find that they are in bijection with the paths on 4 vertices. The following theorem is the base case to which the proof of Theorem 3.4 is reducing.

Theorem 3.7. *For the Bradley–Terry model with four alternatives there are* 20 *saturated designs. Among those*

- 8 *have an empty region of optimality.*
- 12 *have optimal experimental designs.*

The 12 *designs with non-empty region of optimality correspond to the* 12 *labelings of the path P_4. The region of optimality of the path $1-2-3-4$ is constrained by*

$$\begin{aligned} \lambda_{14}(\lambda_{12}+\lambda_{24}) - \lambda_{12}\lambda_{24} &\leq 0, \\ \lambda_{23}(\lambda_{12}+\lambda_{13}) - \lambda_{12}\lambda_{13} &\leq 0, \\ \lambda_{34}(\lambda_{12}\lambda_{24}+\lambda_{12}\lambda_{13}+\lambda_{13}\lambda_{24}) - \lambda_{12}\lambda_{13}\lambda_{24} &\leq 0. \end{aligned} \tag{3.4.1}$$

The regions of optimality for other paths arise from this by relabeling.

Since the D-optimality criterion is invariant under the Sym(4) action by Proposition 3.2, it suffices to study one labeling for each unlabeled graph with three edges on four vertices. The proof of Theorem 3.7 is split into a discussion of information matrices for the three graphs in Fig. 3.1.

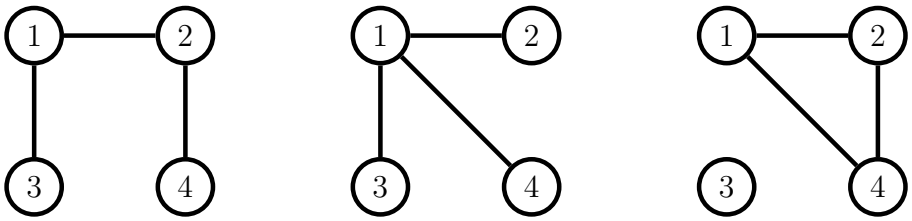


Figure 3.1.: Graph representations of different 3-point designs.

Paths

Consider the path in Fig. 3.1. Its edge set is $\{(1,2),(1,3),(2,4)\}$. A corresponding saturated design can only be optimal if its weights are $w_{12} = w_{13} = w_{24} = \frac{1}{3}$ and $w_{14} = w_{23} = w_{34} = 0$. The information matrix of this design is

$$\begin{aligned} M &= \frac{1}{3}(\lambda_{12} f(1,2) f(1,2)^T + \lambda_{13} f(1,3) f(1,3)^T + \lambda_{24} f(2,4) f(2,4)^T) \\ &= \frac{1}{3} \begin{pmatrix} \lambda_{12}+\lambda_{13} & -\lambda_{12} & -\lambda_{13} \\ -\lambda_{12} & \lambda_{12}+\lambda_{24} & 0 \\ -\lambda_{13} & 0 & \lambda_{13} \end{pmatrix}. \end{aligned}$$

We apply Theorem 2.33. The directional derivatives are

$$g_{ij}(\lambda) := \lambda_{ij} f(i,j)^T M^{-1} f(i,j) - 3.$$

The region of optimality is

$$\{\lambda \in \mathbb{R}^{\mathcal{X}}_{>0} : g_{ij}(\lambda) \leq 0,\ 1 \leq i < j \leq 4\}.$$

This region is a semi-algebraic set (cf. Definition 2.36). Corollary 2.34 simplifies the description because it says that for design points with positive weights the conditions become equations, and those equations have no free variables, as the weights in a saturated design are fixed. Using MATHEMATICA's `Reduce` functionality we derived (3.4.1).

The inequalities in (3.4.1) can be compared to [GS08, Theorem 2]. The structure is similar, but for four alternatives a cubic inequality appears. For m alternatives there are inequality constraints of degree m according to Theorem 3.6. These conditions can be expressed in β-coordinates. The resulting regions of optimality are displayed in Fig. 3.2 on the left.

The claw graph $K_{1,3}$

We now show that the graph in the middle of Fig. 3.1, sometimes known as a *claw*, leads to an empty region of optimality. After symmetry reduction it suffices to show that the design $(12, 13, 14)$ cannot be D-optimal. This design would be optimal in the following region given by the three directional derivatives corresponding to the non-edges $(23, 24, 34)$:

$$\lambda_{23} \leq \frac{\lambda_{12}\lambda_{13}}{\lambda_{12}+\lambda_{13}} \wedge \lambda_{24} \leq \frac{\lambda_{12}\lambda_{14}}{\lambda_{12}+\lambda_{14}} \wedge \lambda_{34} \leq \frac{\lambda_{13}\lambda_{14}}{\lambda_{13}+\lambda_{14}}.$$

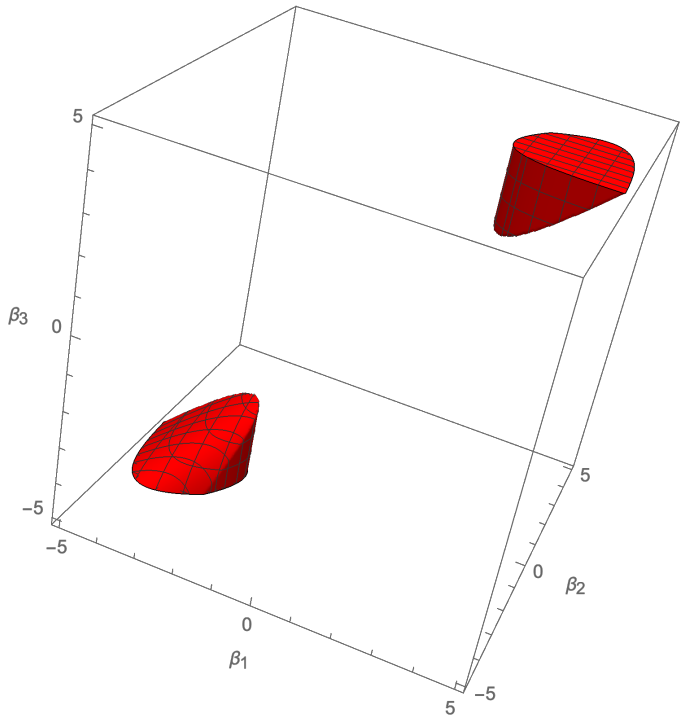


Figure 3.2.: Optimality region for the saturated design on $(12, 13, 24)$.

Plugging in the formulas for the λ_{ij} in terms of the π_i this becomes

$$\begin{aligned}(\pi_2+\pi_3)\left(\pi_1^2+\pi_2\pi_3\right) &\leq \pi_1(\pi_2-\pi_3)^2,\\ (\pi_2+1)\left(\pi_1^2+\pi_2\right) &\leq \pi_1(\pi_2-1)^2,\\ (\pi_3+1)\left(\pi_1^2+\pi_3\right) &\leq \pi_1(\pi_3-1)^2.\end{aligned}$$

Using MATHEMATICA, we find that these conditions are incompatible with the constraints $\pi_1 > 0, \pi_2 > 0, \pi_3 > 0$. An open problem is to find a short certificate for the infeasibility of this system (cf. Section 2.11). An attempt with SOStools [PAV+13] was not successful.

Singular designs

Designs corresponding to the rightmost graph in Fig. 3.1 have singular information matrices and can thereby not be D-optimal, as information matrices are positive semidefinite by definition.

Proof of Theorem 3.7. Since there are 12 distinct labelings of the path on four vertices, the theorem follows from the computations in the subsections on paths, the claw graph and singular designs in Section 3.4.1. □

3.4.2. Unsaturated Designs

We now examine the designs whose support contains at least four pairs. In this case the weights w_{ij} of an optimal design are not necessarily uniform. Instead we find formulas that express the weights in terms of the parameters. These formulas might look complicated, but they are very symmetric and can easily be handled by computer algebra systems. Our approach is again via Theorem 2.33: optimality of a design ξ^* is equivalent to

$$\lambda_{ij} f(i,j)^T M(\xi^*, \beta)^{-1} f(i,j) - 3 \leq 0, \qquad 1 \leq i < j \leq 4. \tag{3.4.2}$$

Furthermore, by Corollary 2.34, there is equality for any pair i, j such that $w_{ij} > 0$ in ξ^*. We distinguish cases according to the size of the support.

Full support

Full support means that all weights of a design are positive. Then all inequalities (3.4.2) hold with equality and we have a system of 6 equations in the variables w_{ij}, λ_{ij} for $1 \leq i < j \leq 4$. We used MATHEMATICA to solve the system and to express the weights w_{ij} as functions of the intensities λ_{ij}:

$$\begin{aligned} w_{ij} = \frac{1}{A}(&\lambda_{ik}\lambda_{il}\lambda_{jk}\lambda_{jl}(\lambda_{ij}\lambda_{ik}\lambda_{il}\lambda_{jk}\lambda_{jl} - \lambda_{ij}\lambda_{ik}\lambda_{il}\lambda_{jk}\lambda_{kl} - \lambda_{ij}\lambda_{ik}\lambda_{il}\lambda_{jl}\lambda_{kl} \\ &- \lambda_{ij}\lambda_{ik}\lambda_{jk}\lambda_{jl}\lambda_{kl} + \lambda_{ij}\lambda_{ik}\lambda_{jk}\lambda_{kl}^2 - \lambda_{ij}\lambda_{ik}\lambda_{jl}\lambda_{kl}^2 - \lambda_{ij}\lambda_{il}\lambda_{jk}\lambda_{jl}\lambda_{kl} - \lambda_{ij}\lambda_{il}\lambda_{jk}\lambda_{kl}^2 \\ &+ \lambda_{ij}\lambda_{il}\lambda_{jl}\lambda_{kl}^2 + 2\lambda_{ik}\lambda_{il}\lambda_{jk}\lambda_{jl}\lambda_{kl})), \end{aligned}$$

where (i, j, k, l) is any permutation of $(1, 2, 3, 4)$. The term A is the normalization that ensures $\sum_{i<j} w_{ij} = 1$. It is therefore invariant under Sym(4) acting on the indices.

$$\begin{aligned} A = 3(&\lambda_{ij}\lambda_{ik}^2\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}^2 + \lambda_{ij}\lambda_{ik}\lambda_{il}^2\lambda_{jk}\lambda_{jl}^2\lambda_{kl}^2 - \lambda_{ij}\lambda_{ik}\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}\lambda_{kl}^2 - \lambda_{ij}^2\lambda_{ik}\lambda_{il}^2\lambda_{jk}\lambda_{jl}\lambda_{kl}^2 \\ &- \lambda_{ij}\lambda_{ik}\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}^2\lambda_{kl} - \lambda_{ij}\lambda_{ik}^2\lambda_{il}^2\lambda_{jk}\lambda_{jl}^2\lambda_{kl} - \lambda_{ij}\lambda_{ik}^2\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}\lambda_{kl} - \lambda_{ij}^2\lambda_{ik}\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}\lambda_{kl} \\ &+ \lambda_{ij}^2\lambda_{ik}^2\lambda_{il}^2\lambda_{jk}\lambda_{jl}\lambda_{kl} - \lambda_{ij}\lambda_{ik}^2\lambda_{il}\lambda_{jk}\lambda_{jl}^2\lambda_{kl}^2 - \lambda_{ij}^2\lambda_{ik}\lambda_{il}\lambda_{jk}\lambda_{jl}^2\lambda_{kl}^2 + \lambda_{ij}\lambda_{ik}^2\lambda_{il}\lambda_{jk}^2\lambda_{jl}\lambda_{kl}^2 \\ &- \lambda_{ij}^2\lambda_{ik}\lambda_{il}\lambda_{jk}^2\lambda_{jl}\lambda_{kl}^2 - \lambda_{ij}^2\lambda_{ik}^2\lambda_{il}\lambda_{jk}\lambda_{jl}\lambda_{kl}^2 - \lambda_{ij}\lambda_{ik}^2\lambda_{il}\lambda_{jk}^2\lambda_{jl}^2\lambda_{kl} + \lambda_{ij}^2\lambda_{ik}\lambda_{il}\lambda_{jk}^2\lambda_{jl}^2\lambda_{kl} \\ &- \lambda_{ij}^2\lambda_{ik}^2\lambda_{il}\lambda_{jk}\lambda_{jl}^2\lambda_{kl} + \lambda_{ij}^2\lambda_{ik}\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}^2\lambda_{kl} + \lambda_{ij}^2\lambda_{ik}^2\lambda_{il}\lambda_{jl}^2\lambda_{kl} + \lambda_{ij}^2\lambda_{ik}^2\lambda_{il}\lambda_{jk}\lambda_{jl}^2\lambda_{kl}^2 \\ &+ \lambda_{ij}^2\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}\lambda_{kl}^2 + \lambda_{ik}^2\lambda_{il}^2\lambda_{jk}^2\lambda_{jl}^2\lambda_{kl}). \end{aligned}$$

This design is locally optimal for some β when $w_{ij} > 0$ for all $1 \leq i < j \leq 4$. Fig. 3.3 shows the optimality region of 6-point-designs.

Example 3.8. A simple example for a design with full support arises when $\beta_i = 0$ for all $1 \leq i \leq 4$. Then $\lambda_{ij} = \frac{1}{4}$ for all $1 \leq i < j \leq n$ and therefore $w_{ij} = \frac{1}{6}$, that is, assigning the same number of repetitions to each comparison, is optimal. Fig. 3.3 and the continuity of the formulas for w_{ij} illustrate that, whenever all β_i are sufficiently small, an optimal design will assign almost equal number of repetitions to each pair (i, j).

Remark 3.9. When working with polynomial equations, Gröbner bases are a powerful tool. The expressions of the w_{ij} in terms of the λ_{ij} can also be found using elimination theory, for example in MACAULAY2 [GS].

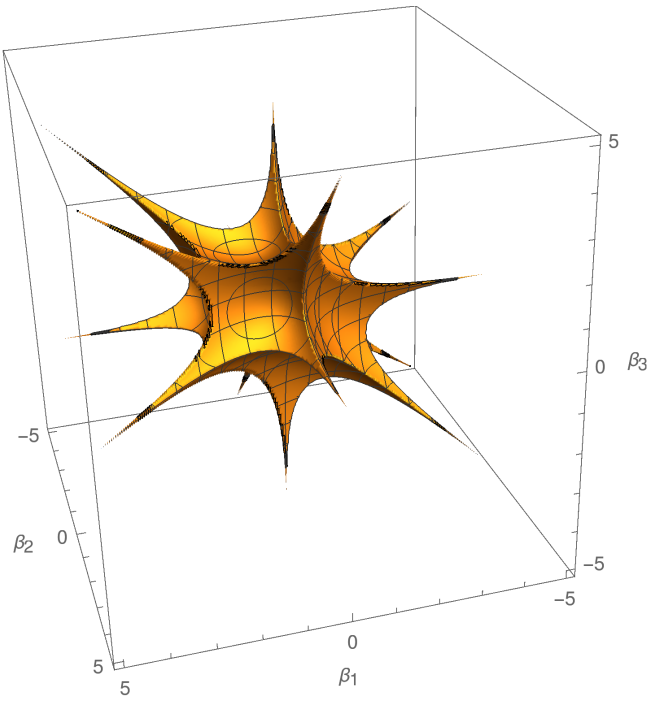


Figure 3.3.: Optimality regions for 6-point designs.

5-point-designs

We now discuss optimal designs where one weight is zero. There is one orbit under the action of Sym(4), that is, a permutation of the alternatives transforms any given five-point-design to the one that does not use comparison $(1, 2)$. Therefore we discuss the design with $w_{12} = 0$ and the remaining weights positive. Then the optimality conditions become

$$
\begin{aligned}
w_{13} &= \frac{2\lambda_{14}\lambda_{34}\left(\lambda_{14}\lambda_{34} - \lambda_{13}\left(\lambda_{14} + \lambda_{34}\right)\right)}{3\left(\lambda_{13}^2\left(\lambda_{14} - \lambda_{34}\right)^2 - 2\lambda_{13}\lambda_{14}\lambda_{34}\left(\lambda_{14} + \lambda_{34}\right) + \lambda_{14}^2\lambda_{34}^2\right)},\\
w_{14} &= \frac{2\lambda_{13}\lambda_{34}\left(\lambda_{13}\left(\lambda_{34} - \lambda_{14}\right) - \lambda_{14}\lambda_{34}\right)}{3\left(\lambda_{13}^2\left(\lambda_{14} - \lambda_{34}\right)^2 - 2\lambda_{13}\lambda_{14}\lambda_{34}\left(\lambda_{14} + \lambda_{34}\right) + \lambda_{14}^2\lambda_{34}^2\right)},\\
w_{23} &= \frac{2\lambda_{24}\lambda_{34}\left(\lambda_{24}\lambda_{34} - \lambda_{23}\left(\lambda_{24} + \lambda_{34}\right)\right)}{3\left(\lambda_{23}^2\left(\lambda_{24} - \lambda_{34}\right)^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}\left(\lambda_{24} + \lambda_{34}\right) + \lambda_{24}^2\lambda_{34}^2\right)},\\
w_{24} &= \frac{2\lambda_{23}\lambda_{34}\left(\lambda_{23}\left(\lambda_{34} - \lambda_{24}\right) - \lambda_{24}\lambda_{34}\right)}{3\left(\lambda_{23}^2\left(\lambda_{24} - \lambda_{34}\right)^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}\left(\lambda_{24} + \lambda_{34}\right) + \lambda_{24}^2\lambda_{34}^2\right)},
\end{aligned}
$$

and with

$$
\begin{aligned}
B = 3\left(\lambda_{13}^2\lambda_{14}^2 - 2\lambda_{13}^2\lambda_{14}\lambda_{34} - 2\lambda_{13}\lambda_{14}\lambda_{34}^2 - 2\lambda_{13}\lambda_{14}^2\lambda_{34} + \lambda_{13}^2\lambda_{34}^2 + \lambda_{14}^2\lambda_{34}^2\right)\\
\cdot\left(\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}^2\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{24}^2\lambda_{34} + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2\right),
\end{aligned}
$$

we obtain

$$w_{34} = \frac{1}{B}\Big(3\lambda_{13}^2\lambda_{14}^2\lambda_{23}^2\lambda_{24}^2 - 4\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}^4 - 2\lambda_{13}\lambda_{14}\lambda_{23}^2\lambda_{24}^2\lambda_{34}^2 + 4\lambda_{13}\lambda_{14}^2\lambda_{23}\lambda_{24}^2\lambda_{34}^2 \\ + 4\lambda_{13}^2\lambda_{14}\lambda_{23}\lambda_{24}^2\lambda_{34}^2 + 4\lambda_{13}\lambda_{14}^2\lambda_{23}^2\lambda_{24}\lambda_{34}^2 + 4\lambda_{13}^2\lambda_{14}\lambda_{23}^2\lambda_{24}\lambda_{34}^2 - 2\lambda_{13}^2\lambda_{14}^2\lambda_{23}\lambda_{24}\lambda_{34}^2 \\ - 4\lambda_{13}\lambda_{14}^2\lambda_{23}^2\lambda_{24}^2\lambda_{34} - 4\lambda_{13}^2\lambda_{14}\lambda_{23}^2\lambda_{24}^2\lambda_{34} - 4\lambda_{13}^2\lambda_{14}^2\lambda_{23}\lambda_{24}^2\lambda_{34} - 4\lambda_{13}^2\lambda_{14}^2\lambda_{23}^2\lambda_{24}\lambda_{34} \\ + 2\lambda_{13}\lambda_{14}\lambda_{23}^2\lambda_{34}^4 + \lambda_{13}^2\lambda_{14}^2\lambda_{23}^2\lambda_{34}^2 + 2\lambda_{13}\lambda_{14}\lambda_{24}^2\lambda_{34}^4 + \lambda_{13}^2\lambda_{14}^2\lambda_{24}^2\lambda_{34}^2 + 2\lambda_{13}^2\lambda_{23}\lambda_{24}\lambda_{34}^4 \\ + \lambda_{13}^2\lambda_{23}^2\lambda_{24}^2\lambda_{34}^2 - \lambda_{13}^2\lambda_{23}^2\lambda_{34}^4 - \lambda_{13}^2\lambda_{24}^2\lambda_{34}^4 + 2\lambda_{14}^2\lambda_{23}\lambda_{24}\lambda_{34}^4 + \lambda_{14}^2\lambda_{23}^2\lambda_{24}^2\lambda_{34}^2 \\ - \lambda_{14}^2\lambda_{23}^2\lambda_{34}^4 - \lambda_{14}^2\lambda_{24}^2\lambda_{34}^4\Big).$$

These designs are optimal if the directional derivative in $(1,2)$-direction is smaller than or equal to zero, which is equivalent to

$$\lambda_{12}(\lambda_{13}(\lambda_{14}(\lambda_{23}(\lambda_{24}-\lambda_{34})-\lambda_{24}\lambda_{34})+\lambda_{34}(\lambda_{23}(\lambda_{34}-\lambda_{24})-\lambda_{24}\lambda_{34})) \\ -\lambda_{14}\lambda_{34}(\lambda_{23}(\lambda_{24}+\lambda_{34})-\lambda_{24}\lambda_{34})) \geq -2\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}.$$

This inequality together with the formulas for the weights and the condition, that all the weights except w_{12} are positive, gives the design region. This region is non-empty. A plot in β-coordinates is in Fig. 3.4.

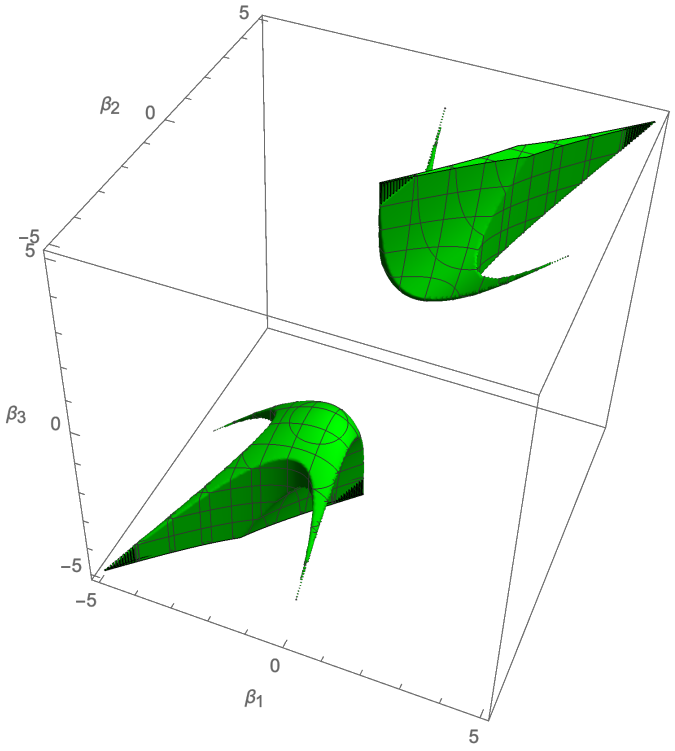


Figure 3.4.: Optimality region of the 5-point-designs with $w_{12} = 0$.

4-point-designs

We now discuss designs whose support contain exactly four points. There are $\binom{6}{4} = 15$ possibilities for such designs which each have two zero weights, $w_{ij} = w_{kl} = 0$. The four point designs form two orbits under the action of Sym(4), distinguished by whether the two non-edges in the graph representation share a vertex or not, that is, whether $|\{i,j,k,l\}| = 4$, that is, i,j,k,l are all distinct, or $|\{i,j,k,l\}| = 3$, that is, exactly two are equal. In the first case, there are three different design classes. We believe that these designs cannot be D-optimal, as the condition $w_{ij} = w_{kl} = 0$ with $|\{i,j,k,l\}| = 4$ implies that a third weight is zero, which would lead to a saturated design. A proof of this statement eludes us so far. Using MATHEMATICA, it follows from the equivalence theorem that such a design satisfies

$$\lambda_{ik}w_{ik}(3w_{ik}-1) = \lambda_{il}w_{il}(3w_{il}-1) = \lambda_{jk}w_{jk}(3w_{jk}-1) = \lambda_{jl}w_{jl}(3w_{jl}-1), \tag{3.4.3}$$

with $w_{\cdot\cdot} < \frac{1}{3}$ for all nonzero weights, and additionally the inequalities

$$\frac{\lambda_{ij}(3(w_{il}+w_{jl})-2)(3(w_{il}+w_{jl})-1)}{\lambda_{jl}w_{jl}(3w_{jl}-1)} \leq 3, \tag{3.4.4}$$

$$\frac{\lambda_{kl}(3(w_{jk}+w_{jl})-2)(3(w_{jk}+w_{jl})-1)}{\lambda_{jl}w_{jl}(3w_{jl}-1)} \leq 3. \tag{3.4.5}$$

Among the solutions of (3.4.3) there are the saturated designs. If one of the weights equals $1/3$, then (3.4.3) implies that another weight is zero, i.e. the design is saturated. Since saturated designs contradict the inequalities (3.4.4) and (3.4.5), we only look for solutions where all of the weights lie in the open interval $(0, 1/3)$. There are solutions of (3.4.3) that satisfy this, for example, if the weights and corresponding intensities are equal. In all the cases we examined, the inequalities (3.4.4) and (3.4.5) are not satisfied.

Problem 3.10. *Show that independent of the λ_{ij}, a simultaneous solution of (3.4.3), (3.4.4), and (3.4.5) is a saturated design.*

Finally we analyze the orbit of four-point-designs assuming $w_{ij} = w_{kl} = 0$ where $|\{i,j,k,l\}| = 3$. Consider the representative with $w_{12} = w_{13} = 0$. Then,

$$\begin{aligned}
w_{14} &= \frac{1}{3},\\
w_{23} &= \frac{2\lambda_{24}\lambda_{34}\left(-\lambda_{23}\lambda_{24}-\lambda_{23}\lambda_{34}+\lambda_{24}\lambda_{34}\right)}{3\left(\lambda_{23}^2\lambda_{24}^2-2\lambda_{23}^2\lambda_{24}\lambda_{34}-2\lambda_{23}\lambda_{24}\lambda_{34}^2-2\lambda_{23}\lambda_{24}^2\lambda_{34}+\lambda_{23}^2\lambda_{34}^2+\lambda_{24}^2\lambda_{34}^2\right)},\\
w_{24} &= \frac{2\lambda_{23}\lambda_{34}\left(-\lambda_{23}\lambda_{24}+\lambda_{23}\lambda_{34}-\lambda_{24}\lambda_{34}\right)}{3\left(\lambda_{23}^2\lambda_{24}^2-2\lambda_{23}^2\lambda_{24}\lambda_{34}-2\lambda_{23}\lambda_{24}\lambda_{34}^2-2\lambda_{23}\lambda_{24}^2\lambda_{34}+\lambda_{23}^2\lambda_{34}^2+\lambda_{24}^2\lambda_{34}^2\right)},\\
w_{34} &= \frac{2\lambda_{23}\lambda_{24}\left(\lambda_{24}\lambda_{24}-\lambda_{23}\lambda_{34}-\lambda_{24}\lambda_{34}\right)}{3\left(\lambda_{23}^2\lambda_{24}^2-2\lambda_{23}^2\lambda_{24}\lambda_{34}-2\lambda_{23}\lambda_{24}\lambda_{34}^2-2\lambda_{23}\lambda_{24}^2\lambda_{34}+\lambda_{23}^2\lambda_{34}^2+\lambda_{24}^2\lambda_{34}^2\right)}.
\end{aligned}$$

This design is optimal if the directional derivatives along $(1,2)$ and $(1,3)$ are smaller than 3, so if

$$\frac{3\lambda_{12}(\lambda_{14}+\lambda_{24})}{\lambda_{14}\lambda_{24}} \leq 3 \ \wedge \ \frac{3\lambda_{13}(\lambda_{14}+\lambda_{34})}{\lambda_{14}\lambda_{34}} \leq 3.$$

This optimality region for this 4-point design is visualized in Fig. 3.5. For each point in the optimality region, the specific weights are computed by the equations above.

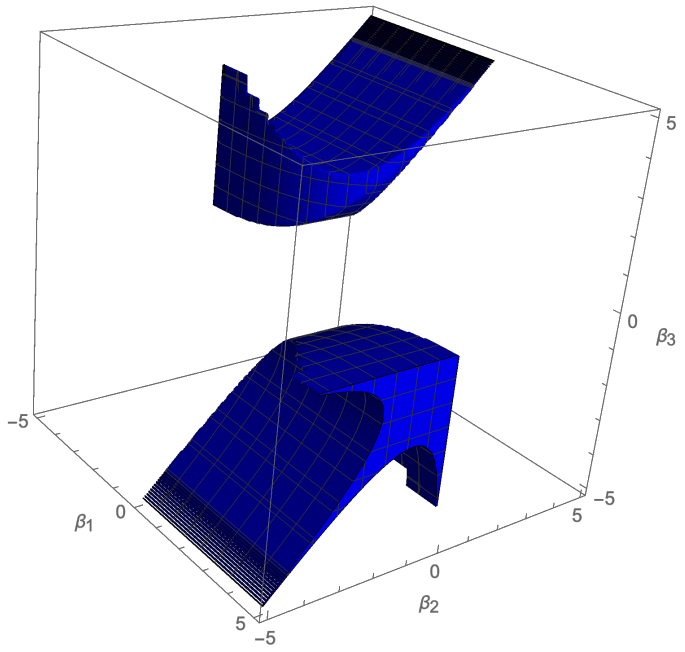


Figure 3.5.: Optimality region for the 4-point design with $w_{12} = w_{13} = 0$.

Having discussed all cases, it suffices to apply the symmetry to each of these regions and then $\mathbb{R}^3$ can be pieced together. Fig. 3.6 gives an idea of this puzzle. The boundaries between the regions always belong to one region. For example, the orange amoeba is open, the red regions for saturated designs are closed (by the non-strict inequalities in Theorem 3.7), and all other regions have both open and closed boundaries.

Remark 3.11. Figures 3.3, 3.4 and 3.5 are reminiscent of the amoebas in tropical geometry. It would be interesting to investigate, if the logarithmic algebraic geometry that arises in β-space from the polynomial constraints in λ-space offers new insights.

3.5. Discussion and outlook

This chapter discussed the parameter regions of optimality for experimental design of the Bradley–Terry model, with the strongest results for 4 alternatives. In practical applications this knowledge can be put to use as follows: First, with a pilot study, initial knowledge of approximate parameters is attained. The initial guess lies in one of the full-dimensional regions illustrated in Fig. 3.6. Depending on which region it is, one can use specific knowledge about the optimal design weights w_{ij}. For example, there are explicit

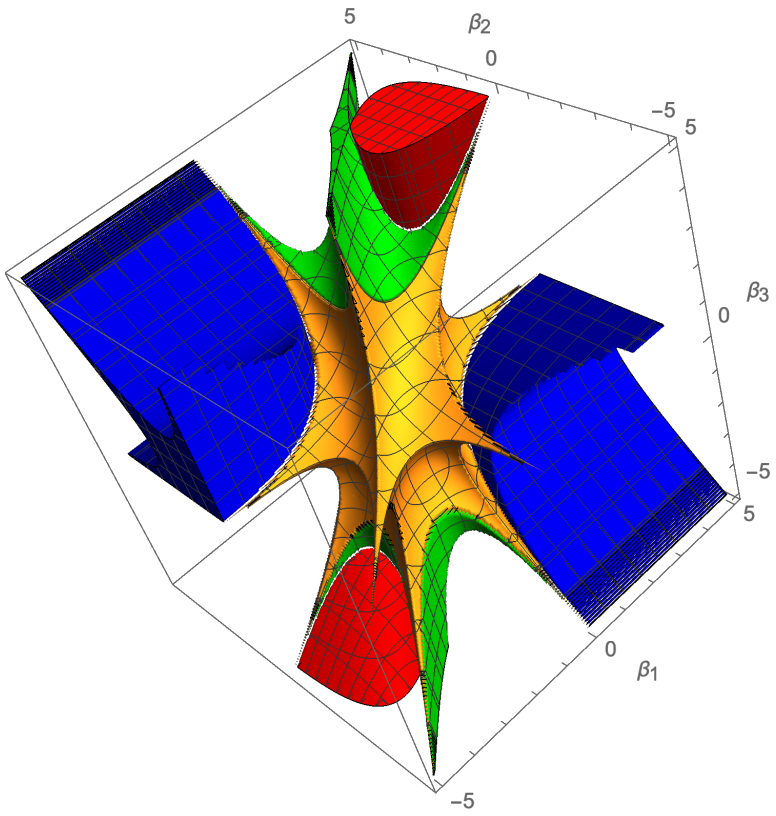


Figure 3.6.: Assembling optimality regions for the Bradley–Terry model.

polynomial formulas for how the optimal weights depend on the location in parameter space. Section 3.4 contains explicit such formulas for the case of 4 alternatives.

In the case that a pilot study reveals parameters in a region where saturated designs are optimal, the solution becomes particularly pleasant: One only needs to assign equal weights to $m-1$ of the pairs. The characterization of regions of optimality of saturated designs is complete, for any number of alternatives (Theorem 3.6).

We illustrate the effect of choosing the right design by computing the efficiency (cf. Eq. (2.8.1)) of the uniform design (assigning equal weights to all pairs) in the case of four alternatives. Consider the line in parameter space that is specified by $2\beta_2 = \beta_1$, $4\beta_3 = 5\beta_1$. Fig. 3.7 shows the efficiency of the uniform design along that line. At $\beta = (\beta_1, \beta_2, \beta_3) = (0, 0, 0)$ the uniform design is optimal. As β grows, the efficiency decreases. First the weights should be adjusted and starting at approximately 1.4 a 5-point design would be optimal. Around 2.1 a 4-point design becomes optimal and finally, from 2.9, a saturated design is optimal. Clearly, working with a uniform design in the case that the support should be smaller is inefficient. In the limit $\beta \to \infty$ the uniform design requires twice as many observations as the optimal saturated design.

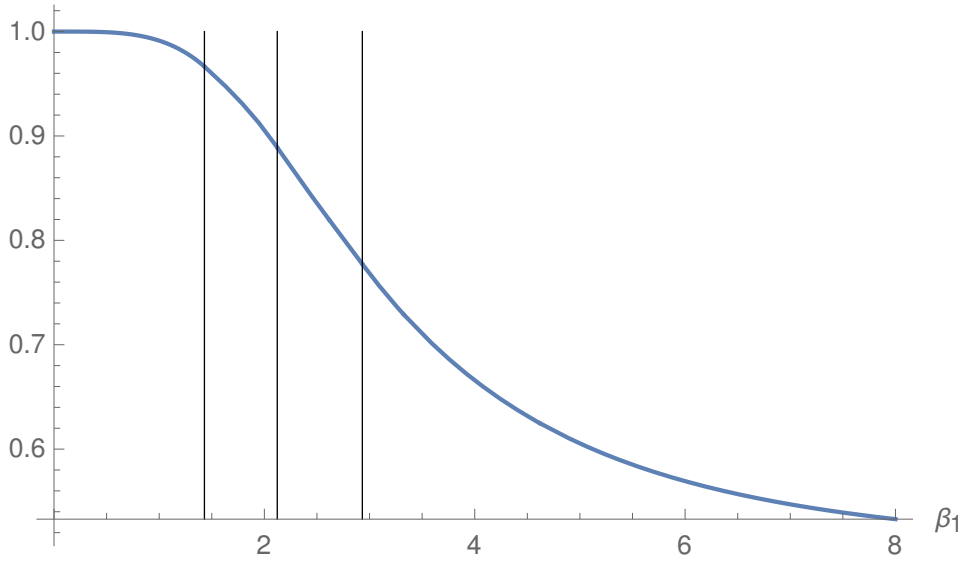


Figure 3.7.: Efficiency of the uniform design along a line in β-space.

3.5.1. Further research directions

We outline some further research directions now. As was pointed out before, the Bradley–Terry model is an example for a generalized linear model. We saw above that if one is interested in precise symbolic solutions for the optimal designs, for example for the D-criterion, it is very useful to apply the Kiefer–Wolfowitz equivalence theorem and study the semi-algebraic geometry of the problem. As explained in Section 2.6, this should apply for a larger class of generalized linear models. Therefore, a natural next step would be an extension of these results to other models, so that afterwards it is possible to formulate a general methodology to find symbolic solutions for locally optimal designs for generalized linear models with a finite design space, see also Section 2.11.

For unsaturated designs, we only discussed the case for four alternatives in this chapter. The most promising and important class of designs suitable for further research are those with full support, as for those by Corollary 2.34 the region of optimality is given by the equations

$$\lambda_{ij} f(i,j)^T M(\xi^*, \beta)^{-1} f(i,j) = (m-1)$$

and positivity constraints $\lambda_{ij} > 0$. We hope that tools from real algebraic geometry can shed further light on such semi-algebraic sets, especially for designs with full support, as their semi-algebraic sets contain no complicated inequalities. We state some observations of the combinatorial structure of full-support designs in Remark 3.13 below.

The Bradley–Terry model considered here is only for the m levels of one attribute and an extension to more attributes is conceivable. The computational challenges of finding optimal designs are formidable and a nice geometry as in the present case is not expected.

In the case of optimality, the equations above express the weights of ξ in terms of the parameters. We conjecture that the equations can be solved in the following sense for

$m \geq 5$.

Conjecture 3.12. *The $\binom{m}{2}$ weights of a fully supported D-optimal design are rational functions in the intensities and of numerator degree $\binom{m}{2} + m - 1$.*

An example of such expressions are the degree 9 equations in Section 3.4.2.

Remark 3.13. The solution for the four-dimensional case reveals that the numerator of a weight w_{ij} is a sum of 10 monomials. These monomials can be described combinatorially as follows. For simplicity, let $i = 1$ and $j = 2$. Then 8 of the 10 monomials are products of the squarefree monomial $\lambda_{12}\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}$ with monomials of the form $\lambda_{ij}\lambda_{ik}\lambda_{kl}$, where (ij, ik, kl) are edges of the 8 graphs that are either paths or trees on four vertices and that do not contain the edge 1–2. Furthermore, the monomials that come from a graph with a node of degree 3 have a positive sign, while the monomials from paths have a negative sign. The remaining two monomials do not show such an easy structure and it remains open, why they are of the form $\lambda_{13}^2\lambda_{14}^2\lambda_{23}^2\lambda_{24}^2(\lambda_{12} + 2\lambda_{34})$. The complete design is generated by permutations acting on the indices of the numerator described above, while the denominator of the weights is just the sum of all the numerators, that is, a normalization.

Remark 3.14. For the case of 3 alternatives, the solutions for the fully-supported design are also rational functions in the intensities, but only of degree 4. The numerator polynomial of w_{12} is $\lambda_{13}\lambda_{23}(\lambda_{13}\lambda_{23} - \lambda_{12}(\lambda_{13} + \lambda_{23}))$. If we multiply this with λ_{12}, we obtain a similar structure as described in Remark 3.13. It is a product of the squarefree polynomial $\lambda_{12}\lambda_{13}\lambda_{23}$ with $\lambda_{ij}\lambda_{ik}$, where (ij, ik) are edges of the 3 graphs that correspond to saturated designs. If the graph contains the edge 1–2, the sign of the polynomial is negative, otherwise it is positive.

From the structure in the case of 4 alternatives, one can at least partially conjecture the structure of a solution in higher dimensions. In the case of 5 alternatives, we conjecture that for full support designs the function that expresses w_{ij} in the intensities λ_{ij} satisfies the following rules: It is quotient of a polynomial divided by a normalization. The numerator polynomial is of degree $\binom{m}{2} + m - 1$ (i.e. 14 for $m = 5$) and composed as follows. Start with the monomial $\lambda_{12}\lambda_{13}\cdots\lambda_{m-1,m}$. To construct the weight for the comparison 1–2, multiply it with a square-free product of $m - 1$ of the variables λ_{ij}, where ij is an edge in a spanning tree on $[m]$ which does not contain 1–2. Sum these monomials over all trees that do not contain 1–2. For n=5, only 50 out of the 125 trees qualify. In this summation, trees of maximal degree 2 receive a negative sign, the others a positive sign. Additionally, we may have to add monomials of a still unknown structure as in Remark 3.13 above. We expect a similar structure in the denominator for 5 alternatives as for four, so that there is a sum of monomials in the denominator that is multiplied with 4. As there are 125 trees, this would make 500 monomials from the tree-structure. This coincides with having 50 monomials from trees in the numerator, as there are 10 weights for 5 alternatives. In comparison, for 4 alternatives, there are $3 \cdot 22 = 66$ monomials in the denominator, but only $6 \cdot 8 = 48$ come from the described graph structure. The implications of these observations are still unknown.

4. Optimality regions for designs in multiple linear regression models with correlated random coefficients

The chapter is structured as follows: We begin with the setup of the model in Section 4.1 and introduce rhombic designs for multiple linear models in Section 4.2. We study optimality regions for rhombic designs in Section 4.3 and present examples in Section 4.4. The chapter ends with a discussion in Section 4.5.

4.1. General setup

We consider a random coefficient regression model $Y_i(x_i) = f(x_i)^T b_i + \varepsilon_i$, $i = 1, \dots n$ for observations Y_i at experimental settings x_i where $f(x)$ is a p-dimensional vector of linearly independent regression functions, b_i is a p-dimensional vector of random coefficients and ε_i are additional observational errors. The random coefficients are assumed to be distributed with unknown mean vector β and prespecified dispersion matrix D, whereas the error terms ε_i are distributed with zero mean and equal variance σ_ε^2. Moreover the random coefficients and the error terms are assumed to be uncorrelated. In this chapter we assume that all observations Y_i are independent, i.e. only one observation is made for each realization b_i of the random coefficients. Furthermore, we assume here that an intercept is included in the model ($f_1(x) \equiv 1$) such that the additive observational error ϵ_i may be subsumed into the random intercept. This can be achieved by substituting the first entry b_{i1} in the random coefficient vector by $b_{i1} + \varepsilon_i$ and the first entry d_0 in the dispersion matrix D by $d_0 + \sigma_\varepsilon^2$. The model can hence be rewritten as a heteroscedastic linear fixed effects model,

$$Y_i(x_i) = f(x_i)^T \beta + \varepsilon_i, \tag{4.1.1}$$

where now $\varepsilon_i = f(x_i)^T(b_i - \beta)$ with mean zero and the variance function defined by $\sigma^2(x) = f(x)^T D f(x)$. Within this heteroscedastic linear model for each single setting x in a design region $\mathcal{X}$ the elemental information matrix [AFHZ14] equals $M(x) = f(x)f(x)^T/\sigma^2(x)$, assuming that $\sigma^2(x) > 0$ for all $x \in \mathcal{X}$. Then for a design ξ as introduced in Section 2.6, the standardized (per observation) information matrix is given by $M(\xi) = \sum_{j=1}^m \xi(x_j) M(x_j)$, which is proportional to the finite sample information matrix with a normalizing constant $\frac{1}{n}$. Note that the covariance matrix for the weighted least squares estimator $\hat{\beta}$, which is the best linear unbiased estimator for β is proportional

to the inverse of the information matrix. Hence, maximizing the information matrix is equivalent to minimizing the covariance matrix of $\hat{\beta}$.

To compare different designs we consider the most popular criterion, the D-criterion, see Section 2.8.1 for an introduction. As discrete optimization on the set of exact designs is generally to complicated we relax the condition on the weights $\xi(x_j)$ being multiples of $\frac{1}{n}$ and consider approximate designs as explained in Section 2.7. In the setting of approximate designs, for which the proportions $\xi(x)$ are not necessarily multiples of $1/n$, where n denotes the sample size, the D-optimality of a design ξ^* can be established by Theorem 2.31. It holds that a design ξ^* is D-optimal on $\mathcal{X}$, if and only if $f(x)^T M(\xi^*)^{-1} f(x)/\sigma^2(x) \leq p$, uniformly in $x \in \mathcal{X}$. When we substitute $\sigma^2(x) = f(x)^T D f(x)$ into this relation and rearrange terms, we define

$$\psi(x;\xi) := f(x)^T (pD - M(\xi)^{-1}) f(x)$$

as the suitably transformed sensitivity function. Now, D-optimality is achieved, if

$$\psi(x;\xi^*) \geq 0 \tag{4.1.2}$$

for all $x \in \mathcal{X}$. Furthermore, it follows from Corollary 2.34 that equality is attained in (4.1.2) for design points in the support of an optimal design ξ^*. Therefore it holds that $\psi(x;\xi^*) = 0$ for all $x \in \mathcal{X}$ with $\xi^*(x) > 0$. For notational convenience we define

$$\Gamma(\xi) := pD - M(\xi)^{-1},$$

such that $\psi(x;\xi) = f(x)^T \Gamma(\xi) f(x)$.

4.2. Multiple linear regression

In the following we consider the situation of a multiple linear regression model with K factors where we have n observations

$$Y_i(x_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i \tag{4.2.1}$$

with $x_i = (x_{i1}, ..., x_{iK})^T \in \mathcal{X} = [-1, 1]^K$ and variance $\mathbb{V}(\varepsilon_i) = \sigma^2(x_i) = f(x_i)^T D f(x_i)$. Here we assume that we can choose the design points from the symmetric standard hypercube. The vector of regression functions is given by $f(x) = (1, x_1, ..., x_K)^T$, such that the model contains an intercept by the first component of f. Note that now $p = K + 1$ and $\beta = (\beta_0, \ldots, \beta_K)^T$.

We assume that the random coefficients $b_{i1}, ..., b_{iK}$ associated with the components $x_1, ..., x_K$ of the regressor are homoscedastic with variance d_1 and equi-correlated with covariance d_2. Furthermore, let the random intercept b_{i0} be uncorrelated with the other random coefficients. To be more precise we consider a $p \times p$-dimensional dispersion matrix

$$D = \begin{pmatrix} d_0 & 0 \\ 0 & D_1 \end{pmatrix}, \tag{4.2.2}$$

where D_1 is a *completely symmetric* $K \times K$-dimensional matrix

$$D_1 = (d_1 - d_2)I_K + d_2 \mathbb{1}_K \mathbb{1}_K^T. \tag{4.2.3}$$

Here, I_k defines the $K \times K$ identity matrix and $\mathbb{1}_K$ the vector of length K where all entries equal 1.

Definition 4.1 (Model cone)**.** We define

$$\mathcal{C}_K := \left\{ (d_0, d_1, d_2)^T \in \mathbb{R}^3 \;\middle|\; d_0 > 0, d_1 > 0, -\frac{d_1}{K-1} \leq d_2 \leq d_1 \right\}$$

as the *model cone*, so the values of $(d_0, d_1, d_2)^T$ where D is positive semidefinite.

4.2.1. Diagonal dispersion matrix

To start we first assume additionally that $d_2 = 0$ which means that all components of the random coefficients are uncorrelated, so that the dispersion matrix

$$D = \begin{pmatrix} d_0 & 0 \\ 0 & d_1 I_K \end{pmatrix}$$

of the random coefficients b_i is a diagonal matrix. Hence, the variance of each design point is equal to $\sigma^2(x) = d_0 + d_1 \sum_{j=1}^K x_j^2$. In [GDHS12] it is shown that uniform full factorial 2^K-designs supported on the points $(\pm x_1, ..., \pm x_K)$ are D-optimal. It holds that $x_1 = \ldots = x_K = x^*$ is optimal and it depends on the values of d_0 and d_1 if $x^* = 1$ or $x^* < 1$. The designs constitute the orbit generated by $(x_1, ..., x_K)$ with respect to the (finite) group of transformations of both sign changes within the factors and permutations of the factors themselves. For more details see [GDHS12].

4.2.2. Non-diagonal dispersion matrix

We now assume that $d_2 \neq 0$. To reduce the complexity of the system of polynomial equations and inequalities given by the equivalence theorem, one can apply various methods. One of these methods is to assume a certain design structure. A standard approach is the assumption of symmetry in the design under some group action and a restriction of the design region. This is motivated from the symmetry of D-optimal designs for the situation with $d_2 = 0$ as described above in Section 4.2.1. This applies as the D-criterion $\log\det(M(\xi))$ is not affected by these transformations g as $\det(M(\xi^g)) = \det(M(\xi))$. Hence, by convexity the class of invariant designs constitutes an essentially complete class such that search may be restricted to invariant designs. A particular class of invariant designs are rhombic designs. Let $\mathrm{Sym}(K)$ denote the permutation group on K elements and $\{\pm 1\}$ the permutation group with respect to a global sign change, which is therefore isomorphic to $\mathrm{Sym}(2)$.

Definition 4.2. Let a design ξ for the given model with support on the space diagonals without the origin and at most two points per space diagonal be a *rhombic design* if it is invariant under the action of $\mathrm{Sym}(K) \times \{\pm 1\}$ on the design points.

This means that we study designs on $[-1,1]^K$, that are invariant under permutations among the entries of each design point and a global sign change with support on the diagonals of maximal length without the origin. There are $\lfloor \frac{K}{2} \rfloor + 1$ different orbits under this group action, where $\lfloor z \rfloor$ denotes the integer part of some $z \in \mathbb{R}$. We define $\tilde{K} := \lfloor \frac{K}{2} \rfloor$. Now, let x_ℓ for $\ell \in \{0, 1, \ldots, \tilde{K}\}$ denote the location parameter of each orbit $\mathcal{O}_\ell(x_\ell)$ with $0 < x_\ell \leq 1$ and either ℓ or $K - \ell$ negative signs. The location parameter denotes the absolute value of the entries of the design points in $\mathcal{O}_\ell(x_\ell)$ as the design points of a rhombic design are restricted to the space diagonals. Let $N_\ell = 2\binom{K}{\ell}$ for $\ell \neq \frac{K}{2}$ and $N_\ell = \binom{K}{\ell}$ for $\ell = \frac{K}{2}$.

Rhombic designs can be characterized as follows: Let x_j for $j \in \{0, 1, \ldots, \tilde{K}\}$ be a design point with entries of the same absolute value, i.e. x_j lies on a space diagonal of $\mathcal{X}$. Let $\mathcal{O}_j$ be the orbit of x_j under the action of $\mathrm{Sym}(K) \times \{\pm 1\}$ and let $\bar{\xi}_j$ be the uniform design on $\mathcal{O}_j$ that assigns the proportion $\frac{1}{N_j}$ to each $x \in \mathcal{O}_j$. If every orbit $\mathcal{O}_j$ is attributed with a weight $w_j \geq 0$ such that $\sum_{j=0}^{\tilde{K}} w_j = 1$, then $\sum_{j=0}^{\tilde{K}} w_j \bar{\xi}_j$ is a rhombic design.

To formalize the invariance considerations, let g denote the group action that generates rhombic designs. Then, there exists a matrix Q_g so that $f(g(x)) = Q_g f(x)$.

Figures 4.1 and 4.2 exemplify the two different rhombic design classes that will be studied separately. This distinction is made on the location of the design points. With rhombic vertex designs we refer to rhombic designs, where the support is restricted to the vertices of the hypercube, while non-vertex designs are allowed to have points on both the vertices and the interior or in the interior only. In Fig. 4.1, the blue points denote the orbit of (x_0, x_0), with $0 < x_0 \leq 1$, while the red points denote the orbit of $(x_1, -x_1)$, with $0 < x_1 \leq 1$. Similarly, in Fig. 4.2, the blue points denote the orbit of (x_0, x_0, x_0), with $0 < x_0 \leq 1$, while the red points denote the orbit of $(x_1, x_1, -x_1)$, with $0 < x_1 \leq 1$. We chose the name *rhombic designs* due to the structure of the design points in Fig. 4.1.

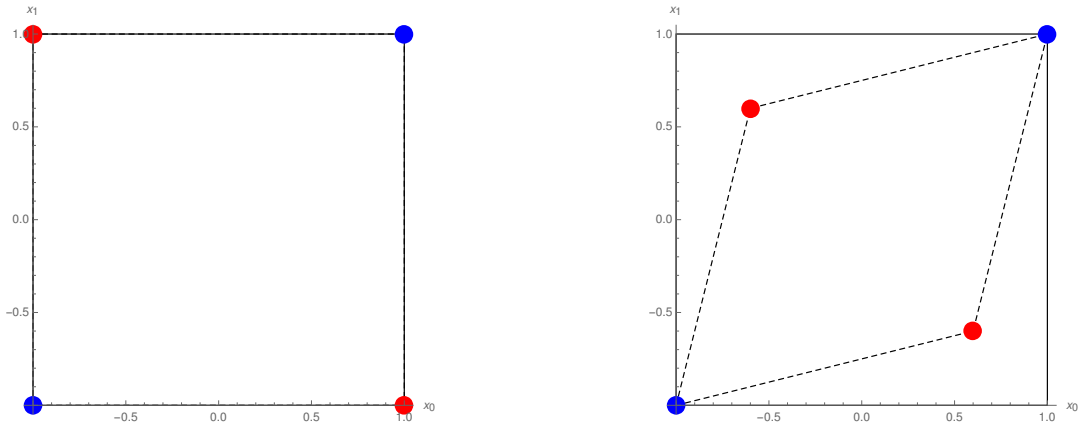
The usefulness of rhombic design is mainly due to the complexity reduction that comes from its definition: Instead of finding $K+1$ design points each with K location variables for the entries and a variable for the design weight, we restrict the problem to $\tilde{K} + 1$ orbits with one weight variable and one location variable per orbit.

Lemma 4.3. *The variance function $\sigma^2(x) = f(x)^T D f(x)$ is equal for all x in one orbit.*

Proof. Let g be the group action from above that generates rhombic designs. By the form of D in (4.2.2), it holds for $\sigma^2(x)$ that

$$\begin{aligned} \sigma^2(g(x)) &= f(g(x))^T D f(g(x)) \\ &= f(x)^T Q_g^T D Q_g f(x). \end{aligned}$$

As $Q_g^T D Q_g = D$, it follows that $\sigma^2(x) = \sigma^2(g(x))$. □



(a) Vertex rhombic design (b) Non-vertex rhombic design

Figure 4.1.: Examples for design points for $K = 2$.

The information matrix of a rhombic design is

$$M(\xi) = \sum_{\ell=0}^{\tilde{K}} \frac{w_\ell}{N_\ell \sigma^2(x_\ell)} \sum_{x \in \mathcal{O}_\ell} f(x) f(x)^T. \tag{4.2.4}$$

To compute the matrix $\sum_{x \in \mathcal{O}_\ell} f(x)f(x)^T$, see that the information matrix is for each orbit structured into a scalar entry in the upper left corner and a $K \times K$ lower right completely symmetric block matrix. This leads to the following Lemma:

Lemma 4.4. *In the setting of Section 4.2, a rhombic design ξ has an information matrix of the form*

$$M(\xi) = \begin{pmatrix} m_0(\xi) & 0 \\ 0 & M_1(\xi) \end{pmatrix},$$

where $m_0(\xi) = \sum_{\ell=0}^{\tilde{K}} \frac{w_\ell}{\sigma^2(x_\ell)}$ and

$$M_1(\xi) = \sum_{\ell=0}^{\tilde{K}} \frac{w_\ell}{N_\ell \sigma^2(x_\ell)} \sum_{x \in \mathcal{O}_\ell} xx^T = (m_1(\xi) - m_2(\xi)) I_K + m_2(\xi) \mathbb{1}_K \mathbb{1}_K^T$$

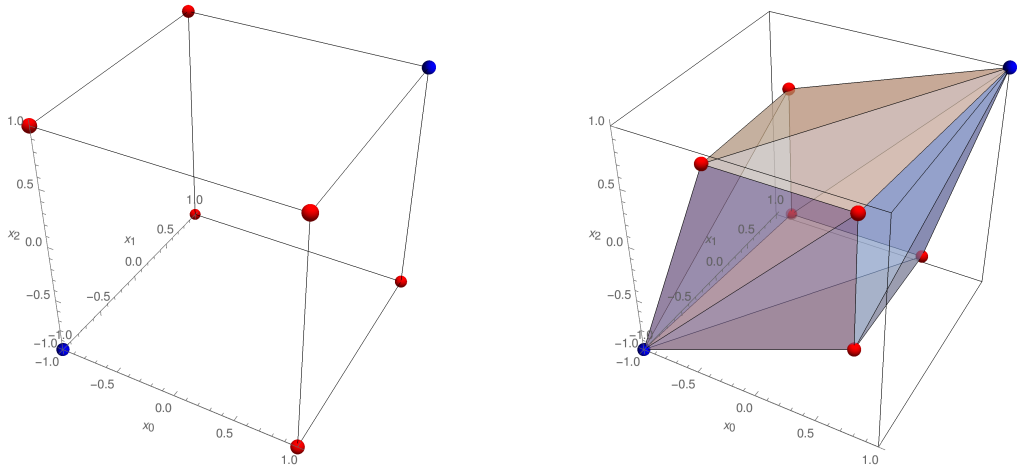
is a completely symmetric $K \times K$ matrix.

Proof. As for every $x \in [-1, 1]^K$, $-x$ lies in the same orbit as x. Therefore, as

$$f(x)f(x)^T + f(-x)f(-x)^T = 2 \begin{pmatrix} 1 & 0 \\ 0 & xx^T \end{pmatrix},$$

$M(\xi)$ is of the form

$$M(\xi) = \begin{pmatrix} m_0(\xi) & 0 \\ 0 & M_1(\xi) \end{pmatrix},$$



(a) Vertex rhombic design (b) Non-vertex rhombic design

Figure 4.2.: Examples for design points for $K = 3$.

where

$$M_1(\xi) = \sum_{\ell=0}^{\tilde{K}} \frac{w_\ell}{N_\ell \sigma^2(x_\ell)} \sum_{x \in \mathcal{O}_\ell} xx^T$$

is a $K \times K$ symmetric matrix and

$$m_0(\xi) = \sum_{\ell=0}^{\tilde{K}} \frac{w_\ell}{\sigma^2(x_\ell)}.$$

Now, $\sum_{x \in \mathcal{O}_\ell} xx^T$ is completely symmetric, because of the permutation invariance of the orbit $\mathcal{O}_\ell$. As the (weighted) sum of completely symmetric matrices is completely symmetric itself, the Lemma follows. □

We defined $\Gamma(\xi)$ as the matrix in the quadratic form in (4.1.2) coming from the equivalence theorem. By Lemma 4.4, denoting the lower right $K \times K$-submatrix of D by D_1,

$$\Gamma(\xi) = \begin{pmatrix} (K+1)d_0 - m_0(\xi)^{-1} & 0 \\ 0 & (K+1)D_1 - M_1(\xi)^{-1} \end{pmatrix}.$$

$\Gamma(\xi)$ has the same block structure as D and $M(\xi)$ with completely symmetric lower right block of dimension $K \times K$, so

$$\Gamma(\xi) = \begin{pmatrix} \gamma_0 & 0 \\ 0 & (\gamma_1 - \gamma_2) I_K + \gamma_2 \mathbb{1}_K \mathbb{1}_K^T \end{pmatrix}$$

with $\gamma_0 = (K+1)d_0 - m_0(\xi)^{-1}$.

We remind the reader of the following fact: The inverse of a completely symmetric $K \times K$ matrix A with

$$A = (a_1 - a_2) I_K + a_2 \mathbb{1}_K \mathbb{1}_K^T$$

is

$$A^{-1} = \frac{1}{a_1 - a_2} I_K - \frac{a_2}{(a_1 - a_2)(a_1 + (K-1)a_2)} \mathbb{1}_K \mathbb{1}_K^T.$$

With these preparatory results, we are able to investigate the optimality of rhombic designs.

4.3. Rhombic Designs and the Equivalence Theorem

4.3.1. Rhombic Vertex Designs

This section studies rhombic designs with all design points on the vertices of the hypercube. We will use the Kiefer–Wolfowitz equivalence theorem to investigate how the D-optimality of a rhombic vertex design depends on D. The investigation leads to the following theorem:

Theorem 4.5. *For the given model 4.2.1 with a dispersion matrix as defined in 4.2.2, let ξ^* be a rhombic vertex design. If either*

1. *K is even or*
2. *K is odd and ξ^* has support not only on $\mathcal{O}_{\tilde{K}}$,*

then ξ^* is D-optimal if and only if the matrix $\Gamma(\xi^*) = pD - M(\xi^*)^{-1}$ is a diagonal matrix

$$\Gamma(\xi^*) = \begin{pmatrix} \gamma_0 & 0 \\ 0 & \gamma_1 I_K \end{pmatrix},$$

with $\gamma_0 \geq 0$ and $\gamma_1 = -\frac{\gamma_0}{K}$.

Proof. " $\Leftarrow$ " Assuming $\gamma_0 \geq 0$, $\gamma_1 = -\frac{\gamma_0}{K}$ and $\gamma_2 = 0$, it follows that

$$\begin{aligned} \psi(x; \xi^*) &= f(x)^T \Gamma(\xi^*) f(x) \\ &= \gamma_0 + \gamma_1 ||x||^2 \\ &= \gamma_0 \left(1 - \frac{||x||^2}{K}\right). \end{aligned}$$

Hence $\psi(x; \xi^*) \geq 0$ for all $x \in [-1, 1]^K$, as $||x||^2 \leq K$. Therefore, the D-optimality follows from Theorem 2.33.
" $\Rightarrow$ " According to (4.1.2) and Theorem 2.33, a design ξ^* is D-optimal if and only if

$$\psi(x; \xi^*) = f(x)^T \Gamma(\xi^*) f(x) \geq 0$$

for all $x \in \mathcal{X}$. Furthermore, by Corollary 2.34, we know that

$$\psi(x; \xi^*) = 0 \tag{4.3.1}$$

for all support points of ξ^*. Now, by (4.3.1), it follows for any design point $x_\ell \in \mathcal{O}_\ell(1)$, that

$$\psi(x_\ell, \xi^*) = \gamma_0 + K\gamma_1 + (K(K-1) - 4\ell(K-\ell))\gamma_2 = 0. \tag{4.3.2}$$

Assuming that the design is supported on at least two different orbits, this directly implies that γ_2 equals zero, as $(K(K-1)-4\ell(K-\ell))$ is strictly monotone for $0 \leq \ell \leq \tilde{K}$.

Now, say that ξ^* is only supported on a single orbit $\mathcal{O}_\ell$ with $0 \leq \ell \leq \tilde{K}$ for even K and $0 \leq \ell < \tilde{K}$ for odd K. If $\ell = 0$, it is easy to see that the information matrix is singular, therefore such a design cannot be D-optimal. The same is true for even K when $\ell = \tilde{K}$ using the same argument as in [FHS20]. It holds that $\psi(x_\ell, \xi^*) = 0$ for all $x_\ell \in \mathcal{O}_\ell(1)$. From the D-optimality of ξ^* it follows that $\psi(x_{\ell-1}, \xi^*) \geq 0$ for $x_{\ell-1} \in \mathcal{O}_{\ell-1}(1)$ and $\psi(x_{\ell+1}, \xi^*) \geq 0$ for $x_{\ell+1} \in \mathcal{O}_{\ell+1}(1)$, so in the orbits with one less or one more negative entry in the vector. If $\gamma_2 \neq 0$, this would imply $\psi(x_{\ell-1}, \xi^*) > \psi(x_\ell, \xi^*) > \psi(x_{\ell+1}, \xi^*)$ or $\psi(x_{\ell+1}, \xi^*) > \psi(x_\ell, \xi^*) > \psi(x_{\ell-1}, \xi^*)$. This contradicts the assumed D-optimality of ξ^*, therefore $\gamma_2 = 0$ holds. It follows from (4.3.2) and $\gamma_2 = 0$ that

$$\gamma_1 = -\frac{\gamma_0}{K}.$$

This implies that

$$\psi(x; \xi^*) = \gamma_0 \left(1 - \frac{||x||^2}{K}\right).$$

and therefore $\gamma_0 \geq 0$ when the design on the vertices is D-optimal. □

Note that $\psi(x, \xi^*) = 0$ for all vertices x of the hypercube as for those it holds that $||x||^2 = K$.

Corollary 4.6. *A rhombic vertex design with support on either at least two orbits or an orbit $\mathcal{O}_\ell(1)$ with $1 \leq \ell < \tilde{K}$ can only be D-optimal when*

$$(d_1 - d_2)(d_1 + (K-1)d_2) - d_0(d_1 + (K-2)d_2) \leq 0.$$

Proof. From Theorem 4.5 we obtain that

$$\gamma_0 \geq 0, \qquad \gamma_1 = -\frac{\gamma_0}{K}, \qquad \gamma_2 = 0$$

is equivalent to the D-optimality of a rhombic vertex design with support on at least two orbits or an orbit $\mathcal{O}_\ell(1)$ with $1 \leq \ell < \tilde{K}$. The equation system $\{\gamma_1 = -\frac{\gamma_0}{K}, \gamma_2 = 0\}$ has two solutions $m_0(\xi)^\pm$ for $m_0(\xi)$ in dependence of d_0, d_1, d_2 and p that we obtained with MATHEMATICA:

$$m_0^\pm(\xi) = \frac{d_0(p+1) + (p-1)\left(d_1 p + d_1 + d_2 p^2 - 3d_2 p\right)}{2p\left(d_0^2 + d_0(p-1)(2d_1 + d_2(p-3)) + (p-1)^2(d_1 - d_2)(d_1 + d_2(p-2))\right)}$$
$$\pm \frac{\sqrt{d_0^2 + 2d_0(d_1(p-1) + d_2 p(p-3)) + p(p-1)^2\left(\frac{d_1^2}{p} + 2d_1 d_2 \frac{p-3}{p-1} + d_2^2 \frac{p^2-5p+8}{p-1}\right)}}{2\frac{p}{p-1}\left(d_0^2 + d_0(p-1)(2d_1 + d_2(p-3)) + (p-1)^2(d_1 - d_2)(d_1 + d_2(p-2))\right)}.$$

Now, with $\gamma_0 = pd_0 - \frac{1}{m_0(\xi)}$ we see that

$$\begin{aligned}
&\Leftrightarrow 0 \geq \left(pd_0 - \frac{1}{m_0^+(\xi)}\right)\left(pd_0 - \frac{1}{m_0^-(\xi)}\right) \\
&\Leftrightarrow 0 \geq p^2 d_0^2 m_0^+(\xi) m_0^-(\xi) - (m_0^+(\xi) + m_0^-(\xi)) + 1 \\
&\Leftrightarrow 0 \geq (d_1 - d_2)(d_1 + (K-1)d_2) - d_0(d_1 + (K-2)d_2).
\end{aligned}$$

To derive the corollary, check that $pd_0 - \frac{1}{m_0^-(\xi)}$ is always negative on $\mathcal{C}_K$, so that

$$\begin{aligned}
&0 \leq pd_0 - \frac{1}{m_0(\xi)} \\
\Leftrightarrow\ &0 \leq pd_0 - \frac{1}{m_0^+(\xi)} \\
\Leftrightarrow\ &0 \geq \left(pd_0 - \frac{1}{m_0^+(\xi)}\right)\left(pd_0 - \frac{1}{m_0^-(\xi)}\right)
\end{aligned}$$

on $\mathcal{C}_K$. This implies the corollary. □

Remark 4.7. Theorem 4.5 gives a semi-algebraic description of the optimality region of rhombic vertex design in the design weights and the coefficients d_0, d_1 and d_2. This means that

$$\left\{\gamma_0 \geq 0, \gamma_1 = -\frac{\gamma_0}{K}, \gamma_2 = 0\right\}$$

can be interpreted as a semi-algebraic set if one takes into account the constraints that ξ is a design and $(d_0, d_1, d_2)^T \in \mathcal{C}_K$. This allows us to obtain symbolic solutions for the design weights $\xi_\ell^* := \xi^*(\mathcal{O}_\ell(1))$ in dependence of the coefficients d_0, d_1 and d_2.

4.3.2. Non-vertex (rhombic) designs

Instead of restricting to rhombic designs we will discuss a broader class of designs, namely designs with a design point in the interior of the hypercube. This design class naturally includes non-vertex rhombic designs.

Theorem 4.8. *A design ξ^* with at least one design point in the interior of the hypercube is D-optimal if and only if $M(\xi^*) = \frac{1}{K+1} D^{-1}$, which means that $\Gamma(\xi^*) = (K+1)D - M(\xi^*)^{-1}$ is zero.*

Proof. " $\Leftarrow$ " $M(\xi^*) = \frac{1}{K+1} D^{-1}$ implies the D-optimality of ξ^* by Theorem 2.33.
" $\Rightarrow$ " By Theorem 2.33 and Corollary 2.34, ξ^* is D-optimal if and only if

$$\psi(x; \xi^*) = f(x)^T((K+1)D - M(\xi^*)^{-1}) f(x) \geq 0$$

for all $x \in [-1, 1]^K$ and

$$\psi(x; \xi^*) = 0,$$

if x is a design point of ξ^*. Now, with $\Gamma(\xi^*) = (K+1)D - M(\xi^*)^{-1}$,

$$\psi(x;\xi^*) = f(x)^T\Gamma(\xi^*)f(x) = \gamma_0 + \gamma_1||x||^2 + 2\gamma_2 \sum_{1\leq\ell<j\leq K} x_\ell x_j$$

is a quadratic polynomial. It holds that

$$\gamma_0 + \gamma_1||x||^2 + 2\gamma_2 \sum_{1\leq\ell<j\leq K} x_\ell x_j = 0 \tag{4.3.3}$$

for design points x of ξ^*. As ξ^* is a non-vertex design there is an interior point x_1 and additionally K further design points $x_2, ...x_{K+1}$ such that $x_1, ..., x_{K+1}$ span $\mathbb{R}^K$ because the information matrix $M(\xi^*)$ is non-singular. Fix the affine subspace generated by x_1 and x_2. It holds that $\psi(x_1,\xi^*) = \psi(x_2,\xi^*) = 0$ and $\psi(\lambda x_1 + (1-\lambda)x_2, \xi^*) \geq 0$ for all $\lambda \in [0,1]$ and additionally for some $\lambda < 0$ because x_1 is an interior point of $[-1,1]^K$. On the affine subspace generated by x_1 and x_2, $\psi(x;\xi^*)$ is a quadratic polynomial in λ. The only quadratic polynomial that is zero on at least two points and non-negative on at least one point on the line segment between these points as well as for at least one point on the line outside this segment is the zero polynomial. Recursively, this can be extended to higher dimensions. Therefore,

$$f(x)^T((K+1)D - M(\xi^*)^{-1})f(x) \geq 0$$

for all $x \in [-1,1]^K$ can only be achieved as an equation, so

$$(K+1)D - M(\xi^*)^{-1} = 0.$$

Hence, the Theorem follows. □

Corollary 4.9. *An invariant design ξ^* with a design point in the interior of the hypercube can only be D-optimal if the first diagonal entry of D^{-1} is larger than the second, which means that*

$$(d_1 - d_2)(d_1 + (K-1)d_2) - d_0(d_1 + (K-2)d_2) > 0.$$

Proof. According to Theorem 4.8, an invariant design ξ^* with a design point in the interior of the hypercube is D-optimal if and only if $\Gamma(\xi^*) = 0$. Now, this implies that $M(\xi^*) = \frac{1}{K+1}D^{-1}$ and therefore

$$m_0(\xi^*) = \frac{1}{(K+1)d_0}, \qquad m_1(\xi^*) = \frac{d_1 + (K-2)d_2}{(K+1)(d_1-d_2)(d_1+(K-1)d_2)},$$

where $m_1(\xi^*)$ denotes the diagonal entries of $M_1(\xi^*)$. It is easy to see that $m_0(\xi^*) > m_1(\xi^*)$ for all designs with interior design points, so we obtain

$$\begin{aligned} & m_0(\xi^*) > m_1(\xi^*) \\ \Leftrightarrow \quad & (d_1 - d_2)(d_1 + (K-1)d_2) - d_0(d_1 + (K-2)d_2) > 0. \end{aligned}$$

□

Remark 4.10. Theorem 4.8 describes the semi-algebraic structure of the optimality area of designs with interior support points. We see that this structure is given by the non-negative real part of an algebraic variety, so the vanishing set of a collection of polynomials under the constraints of the model cone $\mathcal{C}_K$ and the design simplex. This means that we can obtain symbolic solutions for the optimal designs weights and design points in dependence of the coefficients d_0, d_1 and d_2 by studying the set $\{\Gamma(\xi) = 0\}$ under the imposed constraints.

4.4. Rhombic Designs for $K \in \{2, 3, 4, 5\}$

This section investigates for which values d_0, d_1 and d_2 we find a vertex or non-vertex rhombic design for $2 \leq K \leq 5$.

4.4.1. The case $K = 2$

The results of this section where first calculated by hand and later confirmed with a MATHEMATICA implementation of Theorems 4.5 and 4.8. For $K = 2$, it is

$$D = \begin{pmatrix} d_0 & 0 & 0 \\ 0 & d_1 & d_2 \\ 0 & d_2 & d_1 \end{pmatrix} \tag{4.4.1}$$

where $|d_2| < d_1$ and the variance of each design point is equal to

$$\sigma^2(x) = f(x)^T \, D \, f(x) = d_0 + d_1(x_0^2 + x_1^2) + 2d_2 x_0 x_1.$$

The symmetric properties of the covariance structure with respect to the random effects of the two attributes, $\mathbb{V}(b_1) = \mathbb{V}(b_2) = d_1$, motivates us to consider as candidates for the D-optimal designs the following rhombic designs $\xi_{x_0 \lozenge x_1, w}$ consisting of the four design points (x_0, x_0), $(-x_0, -x_0)$, $(-x_1, x_1)$ and $(x_1, -x_1)$ for $x_0, x_1 \in [-1, 1]$ which form a centered rhombus within the design region. Therefore we have $\mathcal{O}_0(x_0) = \{(x_0, x_0), (-x_0, -x_0)\}$ and $\mathcal{O}_1(x_1) = \{(-x_1, x_1), (x_1, -x_1)\}$, so that we obtain $\sigma^2(x_0, x_0) = \sigma^2(-x_0, -x_0)$ and $\sigma^2(-x_1, x_1) = \sigma^2(x_1, -x_1)$. It follows that we can deal with two distinct weights $w_0 = \xi(\mathcal{O}_0(x_0))$ and $w_1 = \xi(\mathcal{O}_1(x_1))$. Since the sum of the weights of all orbits is equal to 1 we can set $w_0 = w$ and $w_1 = 1 - w$. The information matrix for $\xi_{x_0 \lozenge x_1, w}$ results in

$$M(\xi_{x_0 \lozenge x_1, w}) = \begin{pmatrix} \frac{w}{\sigma^2(x_0,x_0)} + \frac{1-w}{\sigma^2(-x_1,x_1)} & 0 & 0 \\ 0 & \frac{w x_0^2}{\sigma^2(x_0,x_0)} + \frac{(1-w)x_1^2}{\sigma^2(-x_1,x_1)} & \frac{w x_0^2}{\sigma^2(x_0,x_0)} - \frac{(1-w)x_1^2}{\sigma^2(-x_1,x_1)} \\ 0 & \frac{w x_0^2}{\sigma^2(x_0,x_0)} - \frac{(1-w)x_1^2}{\sigma^2(-x_1,x_1)} & \frac{w x_0^2}{\sigma^2(x_0,x_0)} + \frac{(1-w)x_1^2}{\sigma^2(-x_1,x_1)} \end{pmatrix}$$

with determinant

$$\det(M(\xi_{x_0 \lozenge x_1, w})) = 4 \frac{(w\,\sigma^2(-x_1, x_1) + (1-w)\,\sigma^2(x_0, x_0))\, w\, (1-w)\, x_0^2\, x_1^2}{(\sigma^2(x_0, x_0)\,\sigma^2(-x_1, x_1))^2}.$$

Maximizing the determinant with respect to the variables x_0, x_1 and w leads to the following results. Note that $d_2 = 0$ is excluded as this was already settled in [GDHS12].

Theorem 4.11. *In the heteroscedastic model of two-factorial multiple regression on* $[-1,1]^2$ *with dispersion matrix* (4.4.1) *it follows:*

(i) If $d_0 \leq d_1 - |d_2|$, *the design* $\xi_{x_0^* \Diamond x_1^*; 0.5}$ *is D-optimal with*

$$x_0^* = \sqrt{\tfrac{d_0}{d_1+d_2}} \quad x_1^* = \sqrt{\tfrac{d_0}{d_1-d_2}}$$

(ii) If $d_1 - |d_2| \leq d_0 \leq \frac{d_1^2-d_2^2}{d_1}$ *the design* $\xi_{x_0^* \Diamond x_1^*, w^*}$ *is D-optimal with*

$$w^* = \frac{2}{3} - \frac{d_0}{6(d_1-d_2)}, \quad x_0^* = \sqrt{\frac{d_1-d_2}{d_1+d_2} \cdot \frac{d_0}{2(d_1-d_2)-d_0}}, \quad x_1^* = 1, \quad \text{if} \quad d_2 > 0,$$

$$w^* = \frac{1}{3} + \frac{d_0}{6(d_1+d_2)}, \quad x_0^* = 1, \quad x_1^* = \sqrt{\frac{d_1+d_2}{d_1-d_2} \cdot \frac{d_0}{2(d_1+d_2)-d_0}}, \quad \text{if} \quad d_2 < 0$$

(iii) If $\frac{d_1^2-d_2^2}{d_1} \leq d_0$ *the vertex design* $\xi_{1 \Diamond 1, w^*}$ *is D-optimal where* w^* *solves the equation*

$$2\left(d_2\left(6w^2 - 6w + 1\right) + d_1(1-2w)\right) + d_0(1-2w) = 0.$$

Proof. Check with Theorems 4.5 and 4.8, that the designs are optimal. □

(i) and (ii) describe non-vertex rhombic designs, while (iii) describes the rhombic design with support on the vertices of the square. The Theorem shows that there is a D-optimal rhombic design for all $(d_0, d_1, d_2)^T \in \mathcal{C}_2$. Figures 4.3 and 4.4 visualize the different optimality regions in Theorem 4.11 for $d_0 = 1$. Note that the region only depends on the quotients $\frac{d_1}{d_0}$ and $\frac{d_2}{d_0}$, so the choice of d_0 is arbitrary.

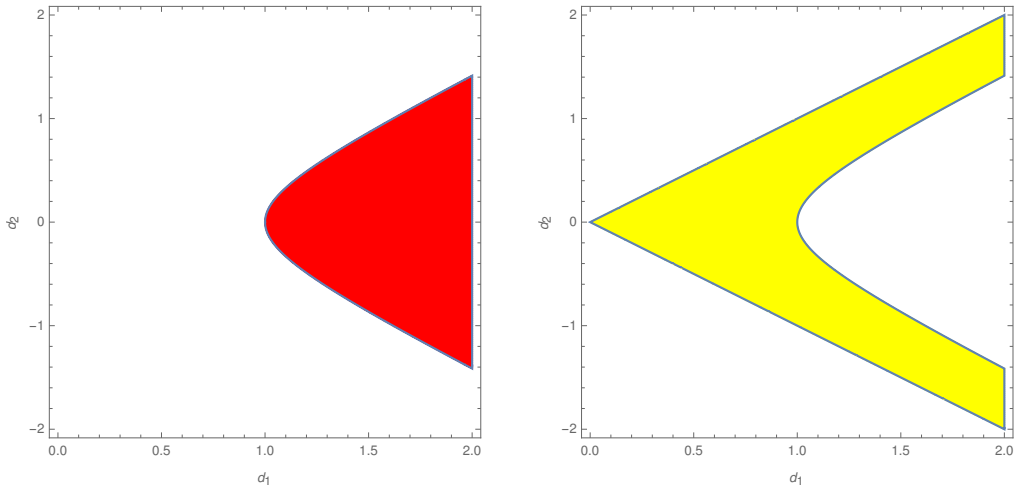
4.4.2. The case $K = 3$

The following Theorem results from a MATHEMATICA implementation of Theorems 4.5 and 4.8.

Theorem 4.12. *For the setting from Section 4.2 with* $K = 3$, *so*

$$D = \begin{pmatrix} d_0 & 0 & 0 & 0 \\ 0 & d_1 & d_2 & d_2 \\ 0 & d_2 & d_1 & d_2 \\ 0 & d_2 & d_2 & d_1 \end{pmatrix},$$

where $-\frac{d_1}{2} < d_2 < d_1$ *it follows:*



(a) Regions (i), (ii) (b) Region (iii)

Figure 4.3.: Parameter regions for $K = 2$: Figure (a) shows parameter regions where rhombic non-vertex designs are D-optimal, while Figure (b) shows parameter regions where rhombic vertex designs are D-optimal.

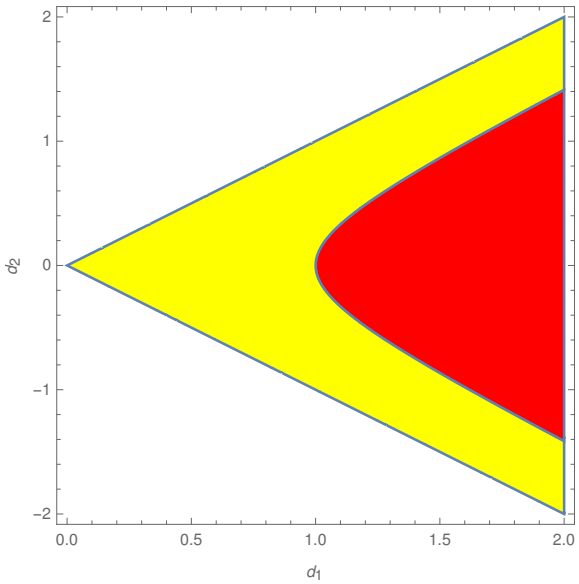


Figure 4.4.: Assembling the parameter regions of rhombic designs for $K = 2$.

(i) *If either* $d_0 < d_1 + d_2 \wedge 0 < d_2 < \frac{d_1}{2}$ *or* $d_0 < \frac{(d_1+2d_2)^2}{d_1-2d_2} \wedge d_2 < 0$, *the design with* $w^* = \frac{1}{4}$ *and*

$$x_0^* = \frac{\sqrt{d_0(d_1 - 2d_2)}}{d_1 + 2d_2}, \qquad x_1^* = \sqrt{\frac{d_0}{d_1 - 2d_2}},$$

is D-optimal.

(ii) *If* $d_2 < \frac{d_1}{2} \wedge d_0 < \frac{(d_1-d_2)(d_1+2d_2)}{d_1+d_2}$, *it holds that the design with*

$$w^* = \frac{3d_0 - 7d_1 + 10d_2}{-16(d_1 - d_2)}, \quad x_0^* = \sqrt{\frac{d_0(-d_1 + 2d_2)}{(d_1 + 2d_2)(3d_0 - 4d_1 + 4d_2)}}, \qquad x_1^* = 1,$$

is D-optimal.

(iii) *If* $d_2 < \frac{d_1}{2} \wedge d_0 < \frac{(d_1-d_2)(d_1+2d_2)}{d_1+d_2}$, *it holds that the design with*

$$x_0^* = 1, \qquad x_1^* = \sqrt{\frac{3d_0(d_1 + 2d_2)}{2d_0d_2 - d_0d_1 - 8d_2^2 + 4d_2d_1 + 4d_1^2}},$$
$$w^* = \frac{(d_1 - 2d_2)(d_0 + 3d_1 + 6d_2)}{16(d_1 - d_2)(d_1 + 2d_2)},$$

is D-optimal.

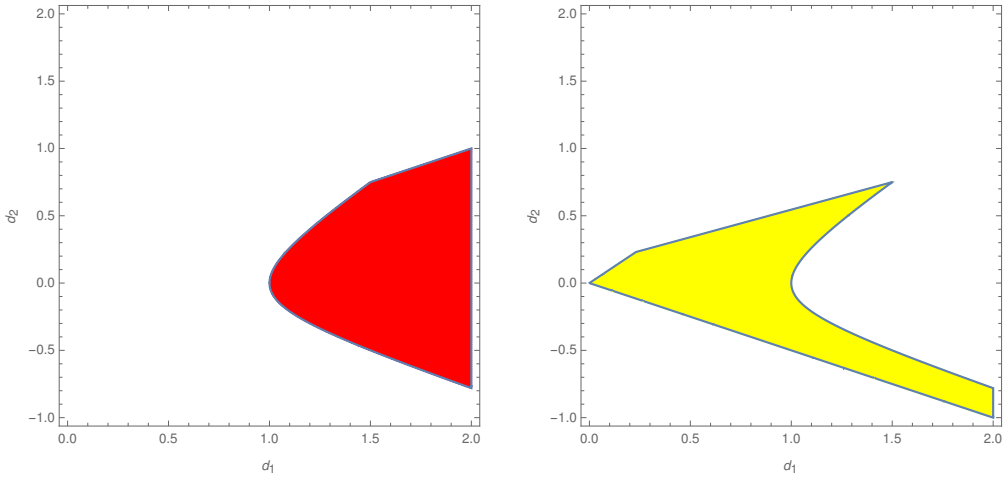
(iv) *If* $d_2 \neq 0$ *and* $\big(d_0(d_1 + d_2) \geq (d_1 - d_2)(d_1 + 2d_2) \wedge \ d_2 \leq \frac{d_1}{2}\big) \vee (\frac{d_1}{2} < d_2 \wedge 3d_0 + 9d_1 > 22d_2)$, *then the design with* $x_0^* = x_1^* = 1$ *and*

$$w^* = \frac{3d_0^2 + 22d_0d_2 + 18d_0d_1 - 120d_2^2 + 66d_2d_1 + 27d_1^2}{64d_2(d_0 - 3d_2 + 3d_1)} - \frac{3\sqrt{(d_0 - 2d_2 + 3d_1)^2\,(d_0^2 + 8d_0d_2 + 6d_0d_1 + 48d_2^2 + 24d_2d_1 + 9d_1^2)}}{64d_2(d_0 - 3d_2 + 3d_1)}$$

is D-optimal.

Proof. For the cases (i), (ii), (iii) check that the equation $\frac{1}{4}D^{-1} = M(\xi^*)$ holds and the model constraints are satisfied. For the fourth case, check that $m_0(\xi^*) \geq \frac{1}{4d_0}$ and that the model constraints are satisfied. □

Note that not all settings of (d_0, d_1, d_2) are covered by Theorem 4.12 and that the described design areas are not disjoint. (ii) and (iii) describe the same optimality area that also contain area (i). Figures 4.5 and 4.6 show the optimality area for $d_0 = 1$ in the (d_1, d_2)-space. Again, the region only depends on the quotients $\frac{d_1}{d_0}$ and $\frac{d_2}{d_0}$, so the choice of d_0 is arbitrary. The area where we did not find an optimal rhombic design is given by $\frac{d_1}{2} < d_2 \wedge 3d_0 + 9d_1 \leq 22d_2$.



(a) Regions (i), (ii), (iii)

(b) Region (iv)

Figure 4.5.: Parameter regions for $K = 3$: Figure (a) shows parameter regions where rhombic non-vertex designs are D-optimal, while Figure (b) shows parameter regions where rhombic vertex designs are D-optimal

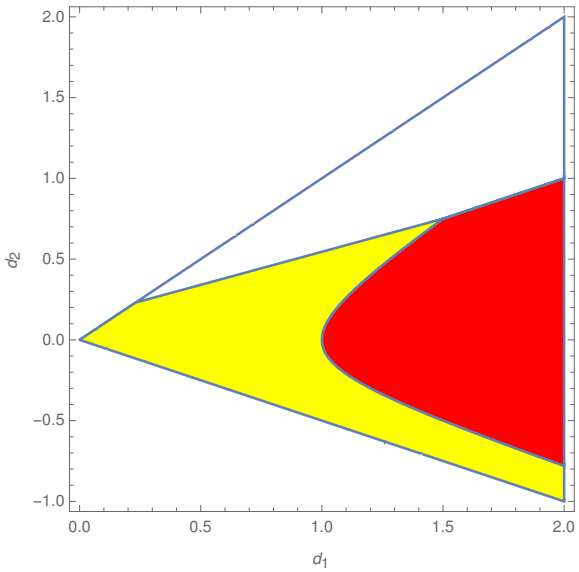
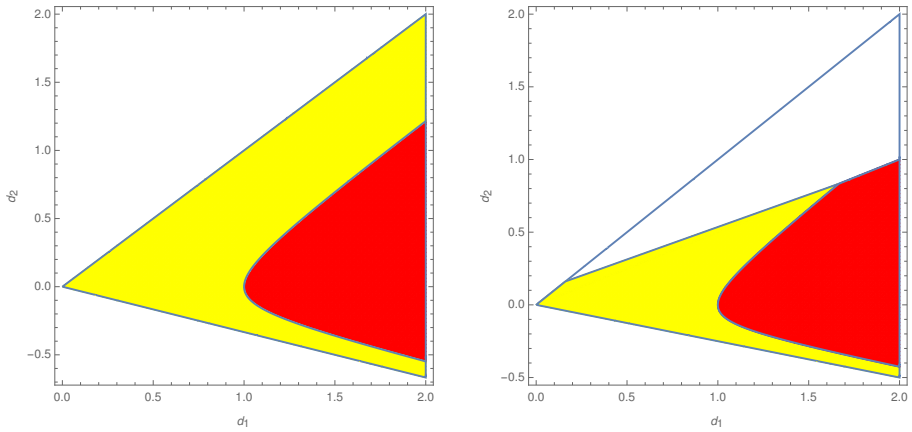


Figure 4.6.: Assembling the parameter regions for $K = 3$.



(a) Parameter regions for $K = 4$. (b) Parameter regions for $K = 5$.

Figure 4.7.: Assembling the optimality regions for $K = 4$ and $K = 5$.

4.4.3. The cases $K = 4$ and $K = 5$

For $K = 4$ and $K = 5$ there are up to three orbits for rhombic designs. To compute an optimal rhombic vertex design, we let $\mathcal{O}_\ell(x_\ell)$ denote the orbits of rhombic design points and choose $x_0 = x_1 = x_2 = 1$, such that the weights are $w_\ell = \xi(\mathcal{O}_\ell(1))$ and check the conditions in Theorem 4.5 for optimality. The different optimality areas are shown in Fig. 4.7. Again, in the red region, a rhombic design with interior points is D-optimal, while in the yellow area, a rhombic vertex design is D-optimal. The separating line is again given by the equality of the first and the second diagonal entry of D^{-1}, see Corollary 4.6 and Corollary 4.9. We see a similar structure as for $K = 2$ and $K = 3$. For $K = 4$, there is a D-optimal rhombic design for every point in $\mathcal{C}_4$, while for $K = 5$, in the region above $d_2 = \frac{d_1}{2}$ there is only a small area where rhombic designs are D-optimal, similar to the case for $K = 3$.

Remark 4.13. The optimality regions shown in the figures for $K \in \{2, 3, 4, 5\}$ are given in the (d_1, d_2)-space while $d_0 = 1$. As before, the region only depends on the quotients $\frac{d_1}{d_0}$ and $\frac{d_2}{d_0}$, so the choice of $d_0 = 1$ is arbitrary. D-optimal designs and the corresponding parameter regions where they are optimal can be found by studying the semi-algebraic sets as described in Remark 4.10 and Remark 4.7. A convenient way to generate the images showing the optimality regions is therefore to use the **Resolve** and **RegionPlot** commands of MATHEMATICA to compute and plot these regions. This was done for $K \in \{2, 3, 4, 5\}$.

4.5. Discussion and outlook

In the preceding sections, optimality regions have been investigated for certain invariant designs in a multiple linear regression model on the hypercube with invariant correlation structure of the random coefficients. It has been shown that for the introduced class of

rhombic designs, it is possible to decide whether a D-optimal design is either supported on the vertices of the hypercube or has interior design points by evaluating a quadratic polynomial depending on the covariance matrix of the random coefficients. This result relies on the Kiefer–Wolfowitz equivalence theorem. The equation separating the two optimality regions is given as the equality of the diagonal entries of D^{-1}.

The results of Theorem 4.8 hold not only for rhombic designs but for all designs with an interior design point, independently of invariance considerations. This means that the D-optimality of designs with interior points is equivalent to the equation $M(\xi^*) = \frac{1}{p}D^{-1}$.

An important observation is the apparent non-existence for D-optimal rhombic designs for certain values of the entries in D. For small dimensions, we have observed that for even K, we could always find a D-optimal rhombic design for any D, while this has not been true for odd K. With respect to our findings, we conjecture the following:

Conjecture 4.14. *For even K, there is a D-optimal rhombic design for all $(d_0, d_1, d_2)^T \in \mathcal{C}_K$. For odd K, there is a D-optimal rhombic design for all $(d_0, d_1, d_2)^T \in \mathcal{C}_K$ with $d_2 \leq \frac{d_1}{2}$.*

In addition to certifying that Conjecture 4.14 is correct, there are several further research directions. A subsequent problem to the conjecture is to show which designs are D-optimal for $d_2 > \frac{d_1}{2}$ for odd K. A more general question is the extension to less restricted covariance structures for the random coefficients. For example, what happens when the covariance matrix of the random coefficients associated with the components $x_1, \ldots, x_K$ is a Hankel matrix (e.g. for an AR(1) process) instead of a completely symmetric matrix?

5. The central limit theorem for a two-sided statistic on Coxeter groups

The structure of the chapter is as follows: Section 5.1 introduces central limit theorems, Wasserstein distances and the little-o-notation. Section 5.2 establishes some basic notations and finite Coxeter groups, so that subsequently, Section 5.3 defines the descent statistic as an example of Coxeter statistics, followed by the descent statistics for the non-dihedral unbounded irreducible types in Section 5.4. The method of interaction graphs, as introduced by Chatterjee [Cha08], is summarized in Section 5.5. We remind the reader of the result of Chatterjee–Diaconis [CD17] in Section 5.6. Section 5.7 and Section 5.8 explain how to generate permutations and signed permutations and their inverses from one vector of random variables, so that the CLT for the statistic $t(\pi) = \mathrm{des}(\pi)+\mathrm{des}(\pi^{-1})$ where π is a random signed permutation follows in Section 5.9. Based on these results, Section 5.10 derives the CLT for this statistic for a sequence of Coxeter groups of type `D`. A generalization to certain statistics of local dependence is presented in Section 5.15.1 and a two-dimensional CLT for the statistic $(\mathrm{des}(\pi), \mathrm{des}(\pi^{-1}))$ in Section 5.15.2. Section 5.11 explains how to recursively derive higher moments of the descent statistic and the statistic t. This is done using conditional expectations and a recursion solver. In Section 5.12, we give sufficient conditions for establishing the CLT for weighted sums of sequences of random variables which all individually satisfy the CLT. Afterwards, we apply the Lindeberg Theorem in Section 5.13 to obtain the asymptotic normality of T_{W_n} for sequences of Coxeter groups W_n which either all are products of dihedral groups or all have only irreducible components of non-dihedral type that satisfy the maximum condition. Combining these results, Section 5.14 delivers the main theorem. In the appendix we present a table of moments of D_{W_n} and T_{W_n} for Coxeter groups of type `A` and `B`.

5.1. Central limit theorems and o-notation

We say that a sequence of integrable random variables $(X_n)_n$ with finite variance *satisfies the central limit theorem (CLT)*, if it holds that

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{D} N(0,1).$$

This means that $(X_n)_n$, normalized by its mean and its standard deviation, converges in distribution towards the standard Gaussian. The following will become useful for establishing CLTs later on:

Lemma 5.1. *Let $(X_n)_n$ be a sequence of integrable random variables with finite variance. Then $(X_n)_n$ satisfies the CLT if and only if every subsequence of $(X_n)_n$ has a subsequence which satisfies the CLT.*

Proof. This follows from the following elementary fact: Let $(a_n)_n$ be a sequence in a topological space A and let $a \in A$. Then if every subsequence of $(a_n)_n$ has a subsequence which converges to a, then $(a_n)_n$ converges to a. □

An approach to show convergence in distribution is to study the Wasserstein distance between a sequence of random variables and some proposed limit. The Wasserstein distance is a distance function on the space of probability measures [AGS05, Chapter 7].

Definition 5.2 (Wasserstein distance, also known as Kantorovich–Rubinstein metric)**.** Let (M, d) be a metric space where every probability measure is a Radon measure and let $P_p(M)$ be the collection of probability measures on M with finite p-th moments. The L^p-Wasserstein distance between $X \sim \mu \in P_p(M)$ and $Y \sim \nu \in P_p(M)$ is defined as

$$\delta_p(\mu, \nu) = (\inf \mathbb{E}\left[d(X, Y)^p\right])^{\frac{1}{p}},$$

where the infimum is taken over all joint distributions of $(X, Y)^T$ on $M \times M$ with marginals μ and ν.

In fact, a bound on the Wasserstein distance that converges towards zero is a stronger result than convergence in distribution. This follows as shrinking Wasserstein distance always implies convergence in distribution, independent of the convergence rate.

We require the definition of the *rank statistic*. Note that this is not related to the group-theoretic notion that is the rank of a Coxeter group, which we will define later.

Definition 5.3. Let $Y = (Y_1, \ldots, Y_n)$ be a vector of real-valued random variables distributed according to a continuous distribution. The rank statistic is defined as $R(Y_i) = \sum_{j=1}^{n} \mathbb{1}_{\{Y_i \geq Y_j\}}$, where $\mathbb{1}_{\{.\}}$ denotes the indicator function. The value of $R(Y_i)$ gives the position of Y_i when Y is sorted in ascending order.

In this work, we use little-o and big-O notation. The definitions vary in the literature, we use the following conventions: Let f and g be maps from $\mathbb{N}_+$ to $\mathbb{R}_{\geq 0}$. We say that $f(n) = o(g(n))$, if it holds that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. Furthermore, we write $f(n) = O(g(n))$, if there is a constant $C > 0$ and $N \in \mathbb{N}$ such that for all $n \geq N$, one has $f(n) \leq Cg(n)$.

5.2. Introduction to finite Coxeter groups

We start with recalling some background about Coxeter groups. For further details, we refer the reader to [BB05].

Let S be a set. A matrix $m : S \times S \to \mathbb{N} \cup \{\infty\}$ is a *Coxeter matrix*, if for all $(s, s') \in S \times S$, the following holds true:

$$m(s, s') = m(s', s) \geq 1,$$
$$m(s, s') = 1 \Leftrightarrow s = s'.$$

A group W is a *Coxeter group*, if there is a set $S \subseteq W$ and a Coxeter matrix $m : S \times S \to \mathbb{N} \cup \{\infty\}$ such that a presentation of W is given by

$$W = \left\langle S \,\middle|\, (ss')^{m(s,s')} = 1 \text{ for all } (s, s') \in S \times S \right\rangle.$$

In this setting, the pair (W, S) is a *Coxeter system* and S the set of *simple reflections*. Instead of representing m with a matrix, one can equivalently display m with a *Coxeter graph*, which is defined with a node set S and edges ss' where $m(s, s') \geq 3$. Hereby edges with $m(s, s') \geq 4$ are labeled with $m(s, s')$. The size of S is the *rank of* (W, S), abbreviated by $\mathrm{rk}(W)$. In what follows, when we talk about a Coxeter group W, we assume that it comes with a fixed generating set S, which makes (W, S) a Coxeter system. Also, if we write W as a product of Coxeter groups $W = W_1 \times W_2 \times \cdots \times W_n$, we assume that $S = S_1 \cup S_2 \cup \ldots \cup S_n$, where S_i is the set of simple reflections of W_i.

A Coxeter group W is *irreducible* if it cannot be written as a non-trivial product of Coxeter groups $W = W_1 \times W_2$. This is equivalent to the corresponding Coxeter graph being connected. By the classification of finite reflection groups (cf. [Cox35]), every *finite* irreducible Coxeter group falls into one of the four infinite families $\mathtt{A}_n$, $\mathtt{B}_n$, $\mathtt{D}_n$, $\mathtt{I}_2(m)$ or is isomorphic to one of seven finite reflection groups of exceptional type. For combinatorial descriptions of the groups of type $\mathtt{A}_n$, $\mathtt{B}_n$, $\mathtt{D}_n$, see [BB05, Chapter 8]. A Coxeter group W is a *dihedral group* or *of dihedral type* if $\mathrm{rk}(W) = 2$. If W is irreducible, this is equivalent to $W \cong \mathtt{I}_2(m)$ for some $m \geq 3$. Any finite Coxeter group W can be written as a product

$$W = W_1 \times W_2 \times \cdots \times W_k,$$

where each W_i is an irreducible Coxeter group. This decomposition is unique up to permutation of the factors and the W_i are the *irreducible components of* W.

The finite reflection groups of infinite type $\mathtt{A}_n$, $\mathtt{B}_n$, $\mathtt{D}_n$ and the dihedral family $\mathtt{I}_2(m)$ are introduced below.

5.2.1. Type $\mathtt{A}_n$

The Coxeter group of type $\mathtt{A}_n$ is generated from the symmetries of the n-simplex. Therefore, the corresponding set S contains all simple reflections in the n-simplex. $\mathtt{A}_n$ is isomorphic to the symmetric group $\mathrm{Sym}(n+1)$, also known as the permutation group. The symmetric group $\mathrm{Sym}(n)$ is isomorphic to the group of all bijective automorphisms on a set $[n] = \{1, \ldots, n\}$. This fact gives rise to the common notation for a permutation as a bijective map $\pi : [n] \to [n]$. It is common to write a permutation $\pi \in \mathrm{Sym}(n)$ in its one-line notation $\pi(1)\pi(2)\ldots\pi(n)$. The Coxeter graph of $\mathtt{A}_n$ is displayed in Fig. 5.1.

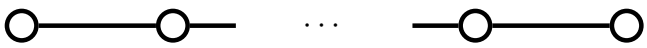


Figure 5.1.: Coxeter graph of type $\mathtt{A}_n$.

5.2.2. Type $\mathtt{B}_n$

The Coxeter group of type $\mathtt{B}_n$ is the symmetry group of the n-hypercube. It is isomorphic to the signed permutation group of rank n, which is the subgroup of all permutations on $\{\pm 1, \ldots, \pm n\}$ with the antisymmetric constraint $-\pi(i) = \pi(-i)$. In a one-line notation we write $\pi = (\pi(1), \ldots, \pi(n))$ where $\pi(i) \in \{\pm 1, \ldots, \pm n\}$ and $\{|\pi(1)|, \ldots, |\pi(n)|\} = [n]$. The Coxeter graph of $\mathtt{B}_n$ is displayed in Fig. 5.2.

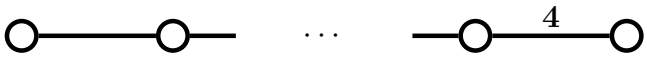


Figure 5.2.: Coxeter graph of type $\mathtt{B}_n$.

5.2.3. Type $\mathtt{D}_n$

The Coxeter group of type $\mathtt{D}_n$ is the symmetry group of the n-demicube. It is isomorphic to the subgroup of the signed permutation group of rank n that consist of all signed permutation with an even number of negative signs. This means, that

$$\mathtt{D}_n = \{\pi \in \mathtt{B}_n : \prod_{i=1}^{n} \pi(i) > 0\}.$$

The Coxeter graph of $\mathtt{D}_n$ is displayed in Fig. 5.3.

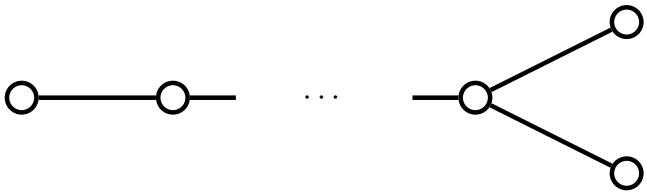


Figure 5.3.: Coxeter graph of type $\mathtt{D}_n$.

5.2.4. Type $\mathtt{I}_2(m)$

The Coxeter group of type $\mathtt{I}_2(m)$ is the symmetry group of the regular m-gon. It is also known as the dihedral group. The dihedral group of rank m has $2m$ elements, as the regular m-gon has m rotational and m reflectional symmetries.The corresponding set S consists of two simple reflections. The Coxeter graph of $\mathtt{I}_2(m)$ is displayed in Fig. 5.4.

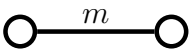


Figure 5.4.: Coxeter graph of type $\mathtt{I}_2(m)$.

5.3. Coxeter statistics

Fix a finite Coxeter group W with a set S of simple reflections. Given an element $w \in W$, the *descent set of* w is defined by

$$\mathrm{Des}(w) \coloneqq \{s \in S \,|\, l_S(ws) < l_S(w)\}\,, \tag{5.3.1}$$

where $l_S(w)$ is the *length of* w *with respect to* S, i.e. the smallest number n such that $w = s_1 s_2 \cdots s_n$, where $s_i \in S$ for all i. The *number of descents* gives rise to a statistic $\mathrm{des} : W \to \mathbb{N}_0$ on W defined by $\mathrm{des}(w) \coloneqq |\mathrm{Des}(w)|$. Choosing an element of W uniformly at random and evaluating this statistic yields a random variable D on $\mathbb{N}$.

The aim of this article is to study the behavior of the statistic t defined by

$$\begin{aligned} t : W &\to \mathbb{N}_0 \\ w &\mapsto \mathrm{des}(w) + \mathrm{des}(w^{-1}). \end{aligned}$$

Just like des, this statistic gives rise to a random variable on $\mathbb{N}$ which is denoted by T. The statistic t was discussed in the case where $W = \mathrm{Sym}(n)$ by Chatterjee–Diaconis [CD17]. They were motivated by the attempt of defining a metric using descents. It also arises in the context of the two-sided analogue of the Coxeter complex recently introduced by Petersen [Pet18].

We also write des_W, D_W, t_W or T_W if we want to emphasize the ambient Coxeter group corresponding to these statistics and random variables.

Lemma 5.4. *Assume that W decomposes as a product $W_1 \times W_2$ of Coxeter groups W_1 and W_2. Then T_W can be written as a sum of independent random variables $T_W = T_{W_1} + T_{W_2}$.*

Proof. Let S_1 and S_2 be the set of simple reflections of W_1 and W_2, respectively. By assumption, we have $S = S_1 \cup S_2$. Every $w \in W$ can be uniquely written as $w = w_1 w_2 = w_2 w_1$, where $w_i \in W_i$ and one has $l_S(w) = l_{S_1}(w_1) + l_{S_2}(w_2)$. Consequently, $\mathrm{des}_W(w) = \mathrm{des}_{W_1}(w_1) + \mathrm{des}_{W_2}(w_2)$ and $t_W(w) = t_{W_1}(w_1) + t_{W_2}(w_2)$. The claim now follows because choosing an element of W uniformly at random is equivalent to choosing uniformly at random w_1 from W_1 and independently w_2 from W_2. □

Theorem 5.5. *Let W be a finite Coxeter group and T as above.*

1. $\mathbb{E}(T) = \mathrm{rk}(W)$.
2. *If W is a product of dihedral groups, $W = \prod_{i=1}^{k} \mathtt{I}_2(m_i)$, then the variance of T is $\mathbb{V}(T) = \sum_{i=1}^{k} \frac{1}{m_i}$.*

3. *If W_n is a sequence of finite Coxeter groups such that for all n, every irreducible component of W_n is of non-dihedral type, then $\mathbb{V}(T_{W_n})$ is of order* $\mathrm{rk}(W_n)$.

Proof. Kahle–Stump computed the variance of T for all types of finite irreducible Coxeter groups in [KS20, Corollary 5.2] . Using Lemma 5.4 and additivity of the variance for independent random variables, the result above follows immediately. □

5.4. Descents on type $\mathtt{A}_n$, $\mathtt{B}_n$ or $\mathtt{D}_n$

As introduced in Section 5.2, elements of the families $\mathtt{A}_n$, $\mathtt{B}_n$ and $\mathtt{D}_n$ allow a one-line notation. Given an element π from either $\mathtt{A}_n$, $\mathtt{B}_n$ or $\mathtt{D}_n$, we now explain how to equivalently describe the set $\mathrm{Des}(\pi)$ in the one-line notation (in comparison to Eq. (5.3.1)). This allows us to obtain the statistic $\mathrm{des}(\pi) = |\mathrm{Des}(\pi)|$ as a sum of indicator functions that depend only on the one-line notation, see also [KS20, Section 2.1]. Subsequently, this shows that the random variable D equals a sum of binary random variables.

5.4.1. Type $\mathtt{A}_{n-1}$

As the Coxeter group of type $\mathtt{A}_{n-1}$ is isomorphic to $\mathrm{Sym}(n)$, the notation for the descents of some permutation $\pi \in \mathrm{Sym}(n)$ simplifies to

$$\mathrm{Des}(\pi) = \{1 \leq i < n : \pi(i) > \pi(i+1)\}.$$

This allows us to write

$$\mathrm{des}_{\mathtt{A}_{n-1}}(\pi) = |\mathrm{Des}(\pi)| = \sum_{i=1}^{n-1} \mathbb{1}_{\{\pi(i)>\pi(i+1)\}}, \tag{5.4.1}$$

where $\mathbb{1}_{\{.\}}$ denotes the indicator function, for the descent statistic.

Example 5.6. Let $\pi = 45213$ be an element in $\mathtt{A}_4$ in one-line notation. We obtain $\mathrm{Des}(\pi) = \{2, 3\}$, as there are descents in the second and third positions in π, as $5 > 2$ and $2 > 1$. Therefore, $\mathrm{des}(\pi) = 2$.

5.4.2. Type $\mathtt{B}_n$

Following [BB05, Proposition 8.1.2], it holds that the descents in some signed permutation $\pi \in \mathtt{B}_n$ in the one-line notation are

$$\mathrm{Des}(\pi) = \{0 \leq i < n : \pi(i) > \pi(i+1)\},$$

where $\pi(0) = 0$. We write for $\pi \in \mathtt{B}_n$

$$\mathrm{des}_{\mathtt{B}_n}(\pi) = |\mathrm{Des}(\pi)| = \mathbb{1}_{\{0>\pi(1)\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{\pi(i)>\pi(i+1)\}}. \tag{5.4.2}$$

Example 5.7. Let $\pi = (4, -5, 2, -1, 3)$ be an element in $\mathtt{B}_5$ in one-line notation. We obtain $\mathrm{Des}(\pi) = \{1, 3\}$, as there are descents in the first and third position in π, as $4 > -5$ and $2 > -1$. Note that there is no descent in the zeroth position, as $\pi(1) = 4 > 0$. Therefore, $\mathrm{des}(\pi) = 2$.

5.4.3. Type $\mathtt{D}_n$

For some $\pi \in \mathtt{D}_n$, it holds that

$$\mathrm{Des}(\pi) = \{0 \leq i < n : \pi(i) > \pi(i+1)\},$$

where $\pi(0) = -\pi(2)$ [BB05, Proposition 8.2.2]. We write for $\pi \in \mathtt{D}_n$

$$\mathrm{des}_{\mathtt{D}_n}(\pi) = |\mathrm{Des}(\pi)| = \mathbb{1}_{\{-\pi(2) > \pi(1)\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{\pi(i) > \pi(i+1)\}}. \tag{5.4.3}$$

Example 5.8. It holds that $\pi = (4, -5, 2, -1, 3)$ is an element of $\mathtt{D}_5$, as $\prod_{i=1}^{5} \pi(i) > 0$. As $-\pi(2) = 5 > 4 = \pi(1)$, it follows that $\mathrm{Des}(\pi) = \{0, 1, 3\}$ and $\mathrm{des}(\pi) = 3$.

5.5. Method of interaction graphs

We give a short overview over the method of interaction graphs as it is presented in [CD17]. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and $f : \mathcal{X}^n \to \mathbb{R}$ a measurable map. Consider a function $G(x)$, which maps every $x \in \mathcal{X}^n$ to a simple graph on $[n] := \{1, 2, \ldots, n\}$. This *graphical rule* is *symmetric*, if for a permutation π the graph $G(x_{\pi(1)}, \ldots, x_{\pi(n)})$ has the edge set

$$\{(\pi(i), \pi(j)) |\ (i, j) \text{ is an edge of } G(x_1, \ldots, x_n)\}.$$

For $m \geq n$, let $G'(x)$ for $x \in \mathcal{X}^m$ be a symmetric graphical rule on $\mathcal{X}^m$. $G'(x)$ is an *extension* of $G(x)$, if $G(x) = G(x_1, \ldots, x_n)$ is the induced subgraph of $G'(x) = G'(x_1, \ldots, x_m)$ for all $x \in \mathcal{X}^m$. To define an interaction rule, let for $x, x' \in \mathcal{X}^n$

$$x^i := (x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n).$$

Furthermore, let x^{ij} be the vector x with replacements in the i-th and j-th position. Then, i and j are *non-interacting* with respect to f, if

$$f(x) - f(x^j) = f(x^i) - f(x^{ij}).$$

A graphical rule G is an *interaction rule* for a function f, if for any $x, x' \in \mathcal{X}^n$ and any i, j, it follows from (i, j) not being an edge of either $G(x), G(x^i), G(x^j)$ or $G(x^{ij})$ that i and j are non-interacting with respect to f. We later apply the following theorem from [Cha08] (see also [CD17, Theorem 4.1]) to signed permutations. The theorem gives a bound on the Wasserstein distance between a normalized statistic that admits a graphical interaction rule and the standard normal distribution.

Theorem 5.9 (Chatterjee). *Let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable map that admits a symmetric interaction rule $G(x)$. Let $X_1, X_2, \ldots$ be independent and identically distributed $\mathcal{X}$-valued random variables and let $X := (X_1, \ldots, X_n)$. Let $F := f(X)$ and $\sigma^2 := \mathbb{V}(F)$. Let $X' = (X'_1, \ldots, X'_n)$ be an independent copy of X. For each j, define*

$$\Delta_j f(X) = F - f(X_1, \ldots, X_{j-1}, X'_j, X_{j+1}, \ldots, X_n)$$

and let $M := \max_j |\Delta_j f(X)|$. Let $G'(x)$ be an extension of $G(x)$ to $\mathcal{X}^{n+4}$ and define

$$\delta := 1 + \text{ degree of the vertex } 1 \text{ in } G'(X_1, \ldots, X_{n+4}).$$

Then the L^1-Wasserstein distance δ_F between $\frac{F-\mathbb{E}(F)}{\sigma}$ and a random variable that is distributed with respect to the standard Gaussian distribution $N(0,1)$ satisfies

$$\delta_F \leq \frac{C\sqrt{n}}{\sigma^2}\mathbb{E}(M^8)^{\frac{1}{4}}\mathbb{E}(\delta^4)^{\frac{1}{4}} + \frac{1}{2\sigma^3}\sum_{j=1}^{n}\mathbb{E}|\Delta_j f(X)|^3$$

for some constant C independent of n.

Chatterjee and Diaconis used the theorem above to show a central limit theorem for statistics of the form $F_1(\pi) + F_2(\pi^{-1})$, where both F_1 and F_2 have bounded local degree and their local components' absolute values are bounded by 1. Hereby π denotes a permutation, hence an element of a Coxeter group of type $\mathtt{A}_n$. We apply the same proof scheme to statistics on signed permutation by modifying their model.

5.6. The CLT for the two-sided descent statistic for type $\mathtt{A}_n$

Given a sequence of Coxeter groups $W_n = \mathtt{A}_n$, so a sequence of permutation groups, the statistic $T_{\mathtt{A}_n}$ satisfies the CLT. This result was first shown in 1996 by Vatutin via generating functions [Vat96]. Chatterjee and Diaconis recently generalized this result in [CD17] to a larger class of locally dependent statistics with a new method that was introduced by Chatterjee in [Cha08]. This *method of interaction graphs* (see Section 5.5) provides a bound on the Wasserstein distance between some normalized statistic that satisfies the method's requirements and the standard normal distribution, see Theorem 5.9. For the statistic $T_{\mathtt{A}_n}$, the following theorem implies the CLT:

Theorem 5.10 (Chatterjee and Diaconis [CD17]). *The Wasserstein distance between $\frac{T_{\mathtt{A}_n}-\mathbb{E}(T_{\mathtt{A}_n})}{\sqrt{\mathbb{V}(T_{\mathtt{A}_n})}}$ and the standard normal distribution is $O\left(n^{-\frac{1}{2}}\right)$.*

The bound in Theorem 5.10 follows from Theorem 5.9 and $\mathbb{V}(T_{\mathtt{A}_n}) = O(n)$.

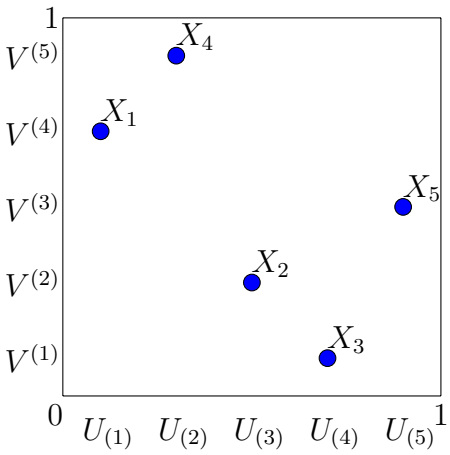


Figure 5.5.: Example for $\pi = 45213$ and $\sigma = 43512$, such that $t(\pi) = 4$.

5.7. Permutations and their inverse from a square

To apply the method of interaction graphs to the statistic

$$T_{\mathtt{A}_{n-1}} = \sum_{i=1}^{n-1} \mathbb{1}_{\{\pi(i)>\pi(i+1)\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{\pi^{-1}(i)>\pi^{-1}(i+1)\}}, \tag{5.7.1}$$

Chatterjee and Diaconis generated both a permutation and its inverse from one set of random variables, namely a vector of uniformly drawn points out of the standard square $[0,1]^2$. Let $\mathcal{X} := [0,1]^2$ and $X_1, X_2, \ldots$ be independent and identically distributed of the form (U_i, V_i) with $(U_i, V_i) \sim \text{Unif}([0,1]^2)$. Let $X := (X_1, \ldots, X_n)$ and let the x-rank of X_i be the rank statistic of U_i (cf. Definition 5.3) among $(U_1, \ldots, U_n)$ and the y-rank of X_i the rank statistic of V_i among $(V_1, \ldots, V_n)$. Then, let $X_{(1)}, \ldots, X_{(n)}$ denote the X_i ordered with respect to their x-ranks and $X^{(1)}, \ldots, X^{(n)}$ with respect to their y-ranks. Let

$$\pi(i) := \text{y-rank of } X_{(i)}, \qquad \sigma(i) := \text{x-rank of } X^{(i)},$$

so that π and σ are random permutations with $\sigma = \pi^{-1}$, as $X^{(i)} = X_{(\sigma(i))}$ and $X_{(i)} = X^{(\pi(i))}$. For an example see Fig. 5.5.

With this construction, Chatterjee and Diaconis were able to apply the method of interaction graphs to prove Theorem 5.10.

5.8. Signed permutations and their inverses from a square

To apply the method of interaction graphs to signed permutations, one has to construct signed permutations similarly to permutations from a set of independent random variables. Now, let $\mathcal{X} := [0,1]^2 \times \{-1, 1\}$ and $X_1, X_2, \ldots$ be independent and identically distributed of the form (U_i, V_i, B_i) with $(U_i, V_i) \sim \text{Unif}([0,1]^2)$ and $B_i \sim \text{Ber}(\frac{1}{2})$ on

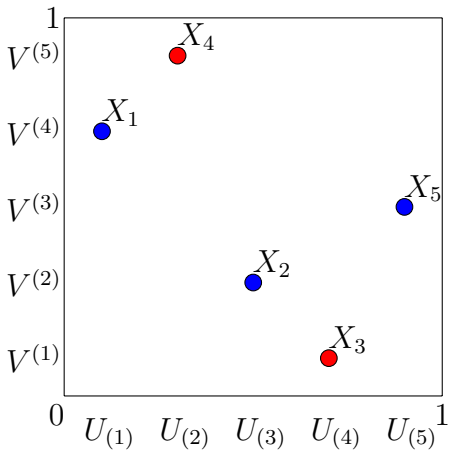


Figure 5.6.: Example for $\pi = (4, -5, 2, -1, 3)$ and $\sigma = (-4, 3, 5, 1, -2)$ with $t(\pi) = 5$. The blue nodes represent a positive sign and the red nodes a negative sign.

$\{-1, 1\}$ and independent of (U_i, V_i). Again, let $X := (X_1, \dots, X_n)$ and let the x-rank of X_i be the rank statistic of U_i among $(U_1, \dots, U_n)$ and the y-rank of X_i the rank statistic of V_i among $(V_1, \dots, V_n)$, so that as before $X_{(1)}, \dots, X_{(n)}$ denote the X_i ordered with respect to their x-ranks and $X^{(1)}, \dots, X^{(n)}$ with respect to their y-ranks. This means that $\pi(i) =$ y-rank of $X_{(i)}$ is a random permutation and $\sigma(i) =$ x-rank of $X^{(i)}$ is its inverse. Now, to see that

$$\tilde{\pi}(i) := B_{(|i|)}\mathrm{sign}(i)\pi(|i|), \qquad \tilde{\sigma}(i) := B^{(|i|)}\mathrm{sign}(i)\sigma(|i|)$$

define random signed permutations, just check that $\tilde{\pi}(-i) = -\tilde{\pi}(i)$ and $\tilde{\sigma}(-i) = -\tilde{\sigma}(i)$ and that $\tilde{\pi}(i)$ and $\tilde{\sigma}(i)$ are injective. Furthermore it follows that $\tilde{\sigma} = \tilde{\pi}^{-1}$, as $B_{(\sigma(|i|))} = B^{(|i|)}$ and

$$\tilde{\pi}(\tilde{\sigma}(i)) = B_{(|\tilde{\sigma}(i)|)}\mathrm{sign}(\tilde{\sigma}(i))\pi(|\tilde{\sigma}(i)|) = B_{(\sigma(|i|))}\mathrm{sign}(B^{(|i|)}\mathrm{sign}(i))\pi(\sigma(|i|)) = i.$$

An example is displayed in Fig. 5.6.

5.9. The CLT for the two-sided descent statistic on signed permutations

As explained above, Chatterjee and Diaconis modeled elements of the symmetric group $\mathrm{Sym}(n)$ and their inverses by ranking functions on series of uniformly distributed random variables on the unit square. We slightly modified this model by additionally introducing a random sign, see Section 5.8. In the following Theorem, we study the asymptotic behavior of the statistic $T_{\mathtt{B}_n}$, which is the sum of the descents in a random signed permutation and its inverse. We show the central limit theorem for said statistic, normalized by its expected value and its variance by adapting the proof of Theorem 1.1 in [CD17] for the modified model.

Theorem 5.11. *Let W_n be a sequence of growing rank of Coxeter groups of type* B*. Then, T_{W_n} satisfies the central limit theorem, if n tends to infinity.*

Proof. As explained in Section 5.8, if π and σ are random permutations with $\pi^{-1} = \sigma$ (cf. Section 5.7), $\tilde{\pi}(i) = B_{(|i|)}\text{sign}(i)\pi(|i|)$ and $\tilde{\sigma}(i) = B^{(|i|)}\text{sign}(i)\sigma(|i|)$ define random signed permutations. It follows that the number of descents in the signed permutation $\tilde{\pi}$ and its inverse $\tilde{\sigma}$ is given by:

$$\begin{aligned} T_{\mathtt{B}_n} := f(X) &= \sum_{i=0}^{n-1} \mathbb{1}_{\{\tilde{\pi}(i)>\tilde{\pi}(i+1)\}} + \sum_{i=0}^{n-1} \mathbb{1}_{\{\tilde{\sigma}(i)>\tilde{\sigma}(i+1)\}} \\ &= \mathbb{1}_{\{0>B_{(1)}\pi(1)\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{B_{(i)}\pi(i)>B_{(i+1)}\pi(i+1)\}} + \mathbb{1}_{\{0>B^{(1)}\sigma(1)\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{B^{(i)}\sigma(i)>B^{(i+1)}\sigma(i+1)\}} \end{aligned}$$

For $x \in \mathcal{X}^n$ with $\mathcal{X} = [0,1]^2 \times \{-1,1\}$ (see Section 5.8), define a simple graph $G(x)$ on $[n]$ as follows: For any $1 \leq i \neq j \leq n$, let $\{i,j\}$ be an edge if and only if the x-rank of x_i and the x-rank of x_j or the y-rank of x_i and the y-rank of x_j differ by at most 1. To check that this graphical rule is symmetric, see that the edge set of a relabeled Graph $G(x_{\pi(1)}, \ldots, x_{\pi(n)})$, where π is an arbitrary permutation, has the edge set $\{(\pi(i), \pi(j))|\ (i,j) \text{ is an edge of } G(x_1, \ldots, x_n)\}$. This is true, since the x-ranks or the y-ranks of $x_{\pi(i)}$ are equal to the respective ranks of x_i. Hence this graph is invariant under relabeling of the indices and it is therefore a symmetric graphical rule. Given $x, x' \in \mathcal{X}^n$, x^i is the vector $(x_1, \ldots x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)$, so the vector x in which the i-th entry is replaced by the i-th entry of x'. Furthermore, x^{ij} is the vector with replacements in the i-th and the j-th entry. Now, suppose that (i,j) is not an edge in $G(x), G(x^i), G(x^j)$ or $G(x^{ij})$. Then, the equation

$$f(x) - f(x^j) = f(x^i) - f(x^{ij})$$

holds, as j is not a neighbor of i in either of the four graphs. To better visualize this, check that

$$f(x) = f(x^i) + f(x^j) - f(x^{ij}). \tag{5.9.1}$$

Any indicator function in $f(x)$, that is not dependent of either x_i or x_j, appears in $f(x^i), f(x^j)$ and $f(x^{ij})$, as it is left unchanged by the replacements in x^i, x^j or x^{ij}. Those indicator functions, that depend on x_i but not on x_j, are unchanged in $f(x^j)$. As i and j are no neighbors in all four graphs, these indicator functions, that depend on x_i but not on x_j, appear in both $f(x^i)$ and $f(x^{ij})$. Therefore, the indicator functions that either depend on x_i or on x_j turn up exactly once on both sides of the equation. Hence Eq. (5.9.1) holds, since there cannot be any indicator functions that depend on both x_i and x_j, as i and j are no neighbors in all four graphs. This means, that $G(x)$ is a symmetric interaction rule for f. Now, we construct an extension $G'(x)$ of $G(x)$ on $\mathcal{X}^{n+4}$. For any $1 \leq i \neq j \leq n+4$, let $\{i,j\}$ be an edge in $G'(x)$ if and only if the x-rank of x_i and the x-rank of x_j or the y-rank of x_i and the y-rank of x_j, differ by at most 5.

As this graph is invariant under relabeling of the indices, it is a symmetric graphical rule. Obviously, every edge in $G(x)$ is also an edge in $G'(x)$, as the distance between two connected nodes in $G(x)$ can be 5 at most through the insertion of four additional nodes. Therefore $G'(x)$ is an extension of $G(x)$. As F and $f(X_1, \dots, X_{j-1}, X'_j, X_{j+1}, \dots, X_n)$ can differ in at most 4 summands, $|\Delta_j f(X)| \leq 4$. Furthermore, the degree of any node in $G'(x)$ is bounded by 20, as either the difference in the x-ranks or in the y-ranks has to be smaller or equal to 5. This means, that $|\delta| \leq 21$. Then, by Theorem 5.9,

$$\delta_{T_{\mathtt{B}_n}} \leq \frac{C\sqrt{n}}{\sigma^2} + \frac{Cn}{\sigma^3}$$

for some constant C. As [KS20] shows, $\sigma^2 = \mathbb{V}(T_{\mathtt{B}_n}) = \frac{n+3}{6}$. Therefore, $T_{\mathtt{B}_n}$ satisfies the central limit theorem.

□

5.10. The CLT for the Coxeter group of type $\mathtt{D}_n$

This section reproduces the previous section's result for elements of Coxeter groups of type $\mathtt{D}_n$. These elements are signed permutations with the constraint to have an even number of negative signs. Therefore, we can reuse the model from the proof of Theorem 5.11, with a slight modification: One sign-generating random variable is set to be the product of all the others. Therefore, the number of negative signs is always even. Of course it is not possible to directly apply the method of interaction graphs, as the local dependency structure is destroyed by one random variable being dependent of all the others. This problem is solved via an application of Slutsky's Theorem.

Theorem 5.12. *Let W_n be a sequence of growing rank of Coxeter groups of type $\mathtt{D}$. Then, T_{W_n} satisfies the central limit theorem, if n tends to infinity.*

Proof. Let $\mathcal{X} := [0,1]^2 \times \{-1,1\}$ and $X_1, X_2, \dots, X_{n-1}$ be independent and identically distributed of the form (U_i, V_i, B_i) with $(U_i, V_i) \sim \mathrm{Unif}([0,1]^2)$ and $B_i \sim \mathrm{Ber}(\frac{1}{2})$ on $\{-1,1\}$. Furthermore, set $X_n = (U_n, V_n, \prod_{i=1}^{n-1} B_i)$ with $(U_n, V_n) \sim \mathrm{Unif}([0,1]^2)$ and $B_n = \prod_{i=1}^{n-1} B_i$. The product of independent $\mathrm{Ber}(\frac{1}{2})$-distributed random variables on $\{-1,1\}$ is again $\mathrm{Ber}(\frac{1}{2})$-distributed on $\{-1,1\}$. Let $X := (X_1, \dots, X_n)$ and let the x-rank and the y-rank of X be defined as in the proof of Theorem 5.11. $X_{(1)}, \dots, X_{(n)}$ denote the X_i ordered in respect to their x-ranks and $X^{(1)}, \dots, X^{(n)}$ in respect to their y-ranks. Then, as in (5.4.3), if $\tilde{\pi} \in \mathtt{D}_n$ and $\tilde{\pi}^{-1} = \tilde{\sigma}$, with $\tilde{\pi}(0) = -\tilde{\pi}(2)$ we obtain

$$\begin{aligned} T_{\mathtt{D}_n} &= \sum_{i=0}^{n-1} \mathbb{1}_{\{\tilde{\pi}(i) > \tilde{\pi}(i+1)\}} + \sum_{i=0}^{n-1} \mathbb{1}_{\{\tilde{\sigma}(i) > \tilde{\sigma}(i+1)\}} \\ &= \mathbb{1}_{\{-B_{(2)}V_{(2)} > B_{(1)}V_{(1)}\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{B_{(i)}V_{(i)} > B_{(i+1)}V_{(i+1)}\}} \\ &\qquad + \mathbb{1}_{\{-B^{(2)}U^{(2)} > B^{(1)}U^{(1)}\}} + \sum_{i=1}^{n-1} \mathbb{1}_{\{B^{(i)}U^{(i)} > B^{(i+1)}U^{(i+1)}\}}. \end{aligned}$$

Now, remove all the indicator functions from $T_{\mathtt{D}_n}$ where $B_{(i)}, B^{(i)}, B_{(i+1)}$ or $B^{(i+1)}$ equal B_n and add indicator functions, so that the resulting random variable is distributed as $T_{\mathtt{B}_{n-1}}$. Then, as $\mathbb{E}(T_{\mathtt{D}_n}) = n$ and $\mathbb{E}(T_{\mathtt{B}_{n-1}}) = n-1$ (see for example in [KS20]),

$$\frac{T_{\mathtt{D}_n} - \mathbb{E}(T_{\mathtt{D}_n})}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}} = \frac{T_{\mathtt{B}_{n-1}} + Y_n - n}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}},$$

where $Y_n = T_{\mathtt{D}_n} - T_{\mathtt{B}_{n-1}}$ is a random variable with $|Y_n| \leq c$ for some positive constant c and all n, so

$$\frac{T_{\mathtt{D}_n} - \mathbb{E}(T_{\mathtt{D}_n})}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}} = \frac{\sqrt{\mathbb{V}(T_{\mathtt{B}_{n-1}})}}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}} \frac{T_{\mathtt{B}_{n-1}} - (n-1)}{\sqrt{\mathbb{V}(T_{\mathtt{B}_{n-1}})}} + \frac{Y_n - 1}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}}. \tag{5.10.1}$$

We know from Theorem 5.11 that $\frac{T_{\mathtt{B}_{n-1}} - (n-1)}{\sqrt{\mathbb{V}(T_{\mathtt{B}_{n-1}})}}$ converges in distribution to a standard normal distribution. Y_n is bounded, as it is a finite sum of indicator functions. Therefore, $\lim_{n\to\infty} \frac{Y_n - 1}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}} = 0$ almost surely and $\lim_{n\to\infty} \frac{\sqrt{\mathbb{V}(T_{\mathtt{B}_{n-1}})}}{\sqrt{\mathbb{V}(T_{\mathtt{D}_n})}} = 1$ (compare [KS20, Corollary 5.2]). Therefore, $T_{\mathtt{D}_n}$ satisfies the central limit theorem (see Slutsky's theorem, for example in [Leh98, Theorem 2.3.3]).

□

5.11. Fourth moments of T

As defined in Section 5.3, let D_W be the random variable associated to the statistic des_W and let T_W be the random variable associated to the statistic t_W for a finite Coxeter group W. The aim of this section is to prove the following theorem:

Theorem 5.13. *Let W be an irreducible Coxeter group of type $\mathtt{A}_n, \mathtt{B}_n$ or $\mathtt{D}_n$. Then the fourth centered moment $\mathbb{E}\left((T_W - \mathbb{E}(T_W))^4\right)$ of T_W is of order n^2.*

In order to show this, we follow and extend the ideas of Özdemir. In [Ö19], he formulated the recursive formulas

$$\mathbb{E}\left(D_{\mathtt{A}_{n+1}} | D_{\mathtt{A}_n}\right) = D_{\mathtt{A}_n} \frac{D_{\mathtt{A}_n} + 1}{n+2} + (D_{\mathtt{A}_n} + 1) \frac{n + 1 - D_{\mathtt{A}_n}}{n+2} = \frac{n+1}{n+2} D_{\mathtt{A}_n} + \frac{n+1}{n+2} \tag{5.11.1}$$

and

$$\mathbb{E}\left(D_{\mathtt{B}_{n+1}} | D_{\mathtt{B}_n}\right) = D_{\mathtt{B}_n} \frac{2D_{\mathtt{B}_n} + 1}{2n+2} + (D_{\mathtt{B}_n} + 1) \frac{n + 1 - 2D_{\mathtt{B}_n}}{2n+2} = \frac{2n-1}{2n+2} D_{\mathtt{B}_n} + \frac{2n+1}{2n+2}. \tag{5.11.2}$$

Here $\mathbb{E}(X|Y)$ denotes the conditional expected value where X is conditioned on Y. Özdemir used these formulas to compute higher moments of $D_{\mathtt{A}_n}$ and $D_{\mathtt{B}_n}$. An important tool for his computations is the smoothing theorem (also known as the the law of total expectation) which can be stated as follows:

Theorem 5.14 (Smoothing Theorem, cf. [Bil95, Theorem 34.4]). *Let X and Y be integrable random variables. Then, it holds that*

$$\mathbb{E}\left(\mathbb{E}\left(X|Y\right)\right) = \mathbb{E}\left(X\right).$$

Our approach for proving Theorem 5.13 is to inductively compute higher moments of T_W and D_W for the different families of Coxeter groups separately. We start in Section 5.11.1 by computing the fourth centered moment of D_W in the case where W is irreducible and of type A or B. These computations serve as an illustration of the methods we use and the results will be needed for our inductive method of computing the fourth centered moments of T_W later on. Building on this, we prove Theorem 5.13 for W of type A and B in Section 5.11.2 and Section 5.11.3, respectively. We finish the proof in Section 5.11.4.

5.11.1. Fourth moment of D

Özdemir showed that the fourth centered moment of the random variable $D_{\mathtt{A}_n}$ is of order n^2 [Ö19, p. 3]. Using the **RSolve** function of MATHEMATICA, we are able to give an explicit formula for this moment:

Lemma 5.15. *Let D_n be the random variable associated to the statistic* des *on the Coxeter group $\mathtt{A}_n$, $n \geq 3$. Then we have:*

$$\mathbb{E}\left(\left(D_n - \mathbb{E}\left(D_n\right)\right)^4\right) = \frac{1}{240}\left(n+2\right)\left(5n+8\right).$$

Proof. From Eq. (5.11.1), we derive the recursion formula

$$\mathbb{E}\left(\left(D_{n+1} - \mathbb{E}\left(D_{n+1}\right)\right)^4 | D_n\right) = \frac{\left(n-2\right)\left(D_n - \mathbb{E}\left(D_n\right)\right)^4}{n+2} + \frac{\left(3n+4\right)\left(D_n - \mathbb{E}\left(D_n\right)\right)^2}{2\left(n+2\right)} + \frac{1}{16}. \tag{5.11.3}$$

By applying $\mathbb{E}$ on both sides of Eq. (5.11.3), the smoothing theorem leads to

$$\mathbb{E}\left(\left(D_{n+1} - \mathbb{E}\left(D_{n+1}\right)\right)^4\right) = \frac{\left(n-2\right)\mathbb{E}\left(\left(D_n - \mathbb{E}\left(D_n\right)\right)^4\right)}{n+2} + \frac{\left(3n+4\right)\mathbb{V}\left(D_n\right)}{2\left(n+2\right)} + \frac{1}{16}$$

and with the formula for the variance found for example in [KS20, Corollary 5.2], we obtain a recursive formula for $a[n] = \mathbb{E}\left(\left(D_n - \mathbb{E}\left(D_n\right)\right)^4\right)$:

$$a[n+1] = \frac{\left(6n+11\right)}{48} + \frac{\left(n-2\right)a[n]}{n+2},$$

which was solved by computing the value $a[3] = \frac{23}{48}$ with SAGE and using the **RSolve** function of MATHEMATICA. □

Using the same method and Eq. (5.11.2), we compute the same moment in type B:

Lemma 5.16. *Let D_n be the random variable associated to the statistic* des *on the Coxeter group $\mathtt{B}_n$, $n \geq 4$. Then we have:*

$$\mathbb{E}\left((D_n - \mathbb{E}(D_n))^4\right) = \frac{1}{240}(n+1)(5n+3). \tag{5.11.4}$$

Proof. From Eq. (5.11.2), we derive the recursion formula

$$\mathbb{E}\left((D_{n+1} - \mathbb{E}(D_{n+1}))^4 | D_n\right) = \frac{(n-3)(D_n - \mathbb{E}(D_n))^4}{n+1} + \frac{(3n+1)(D_n - \mathbb{E}(D_n))^2}{2(n+1)} + \frac{1}{16}. \tag{5.11.5}$$

This is the same recursion formula as for type A_{n-1} in Eq. (5.11.3), so we obtain a recursive formula for $a[n] = \mathbb{E}\left((D_n - \mathbb{E}(D_n))^4\right)$:

$$a[n+1] = \frac{(6n+5)}{48} + \frac{(n-3)\,a[n]}{n+1}, \tag{5.11.6}$$

which was also solved by computing the starting value $a[4] = \frac{23}{48}$ with SAGE and using the **RSolve** function of MATHEMATICA. □

5.11.2. Moments of T for type $\mathtt{A}_n$

Throughout this subsection, let $T_n = T_{\mathtt{A}_n}$, $D_n = D_{\mathtt{A}_n}$ and let D'_n be the random variable associated to the statistic

$$\begin{aligned} \mathtt{A}_n &\to \mathbb{N} \\ w &\mapsto \mathrm{des}\left(w^{-1}\right). \end{aligned}$$

Clearly, we have $T_n = D_n + D'_n$, but D_n and D'_n are not independent. In order to compute the fourth centered moment of T_n, we want to inductively determine mixed moments of the form $\mathbb{E}\left(D_n^k {D'_n}^l\right)$. To compute these moments recursively, we use the following two-dimensional conditional expectation for (D_n, D'_n) introduced by Özdemir:

Lemma 5.17 (see [Ö19, p. 18])**.** *In type $\mathtt{A}_n$, the random variable (D_n, D'_n) satisfies the following:*

$$\mathbb{E}((D_{n+1}, D'_{n+1})|(D_n, D'_n)) = \begin{cases} (D_n, D'_n) & \text{with prob. } P_1 = \frac{(D_n+1)(D'_n+1)+n+1}{(n+2)^2}, \\ (D_n+1, D'_n) & \text{with prob. } P_2 = \frac{(n+1-D_n)(D'_n+1)-n-1}{(n+2)^2}, \\ (D_n, D'_n+1) & \text{with prob. } P_3 = \frac{(D_n+1)(n+1-D'_n)-n-1}{(n+2)^2}, \\ (D_n+1, D'_n+1) & \text{with prob. } P_4 = \frac{(n+1-D_n)(n+1-D'_n)+n+1}{(n+2)^2}. \end{cases}$$

We remark that in comparison to this, there is a shift of indices in [Ö19, p. 18] as there, D_n corresponds to the descent statistic on $\mathrm{Sym}(n) = \mathtt{A}_{n-1}$. Özdemir used this conditional expectation in order to compute the asymptotics of $\mathbb{E}\big((D_n - \mathbb{E}(D_n))^2 (D'_n - \mathbb{E}(D'_n))^2\big)$, see [Ö19, Lemma 5.1]. We obtain his results and generalizations of it in the proof of the following proposition.

Proposition 5.18. *In type* $\mathtt{A}_n$, $n \geq 3$, *the fourth centered moment of* T_n *is given by*

$$\mathbb{E}\left((T_n - \mathbb{E}(T_n))^4\right) = \frac{1}{60}\left(5n^2 + 79n + 258\right) - \frac{5n+2}{n(n+1)}.$$

Proof. Define $U_n := D_n - \mathbb{E}(D_n) = D_n - n$ and $U'_n := D'_n - \mathbb{E}(D'_n)$. Our goal is to compute

$$\mathbb{E}\left((T_n - \mathbb{E}(T_n))^4\right) = \mathbb{E}\left((U_n + U'_n)^4\right).$$

Multiplying out the right hand side of this equation and using linearity of the expected value, we see that it suffices to compute $\mathbb{E}(U_n^k {U'}_n^l)$ for all $0 \leq k, l \leq 4$ with $k + l = 4$. Using the smoothing theorem and Lemma 5.17, we derive the following recursion formula for fixed k and l:

$$\mathbb{E}({U}_{n+1}^k {U'}_{n+1}^l) = \mathbb{E}(U_n^k {U'_n}^l P_1 + (U_n+1)^k {U'_n}^l P_2 + U_n^k (U'_n+1)^l P_3 + (U_n+1)^k (U'_n+1)^l P_4),$$

where P_1, P_2, P_3 and P_4 are as in Lemma 5.17. The right hand side of this equation only depends on $\mathbb{E}(U_n^i {U'_n}^j)$ with $i \leq k$ and $j \leq l$. Hence, inductively computing $\mathbb{E}(U_n^i {U'_n}^j)$ for all pairs (i, j) with $i \leq k$, $j \leq l$ and where at least one of this inequalities is strict, we obtain a recursion formula for $\mathbb{E}(U_n^k {U'_n}^l)$.

To obtain the claimed result, we computed the starting values with Sage and solved the recursion with the **RSolve** command of Mathematica, just as in Section 5.11.1. The intermediate results of these computations can be found in Appendix A.1. □

5.11.3. Moments of T for type $\mathtt{B}_n$

We now turn to type $\mathtt{B}_n$. Let $D_n := D_{\mathtt{B}_n}$, $T_n := T_{\mathtt{B}_n}$ and let D'_n be the random variable associated to

$$\begin{aligned} \mathtt{B}_n &\to \mathbb{N} \\ w &\mapsto \operatorname{des}\left(w^{-1}\right). \end{aligned}$$

To compute the fourth centered moment of $T_n = D_n + D'_n$, we want to take the same approach as in Section 5.11.2. For this, we first need an analogue of Lemma 5.17. We start by setting

$$B_{n,i,j} := \left|\{w \in \mathtt{B}_n \,\middle|\, \operatorname{des}(w) = i \text{ and } \operatorname{des}\left(w^{-1}\right) = j\}\right|.$$

These numbers are the coefficients of the *type* $\mathtt{B}_n$ *two-sided Eulerian polynomial*

$$B_n(s,t) := \sum_{w \in \mathtt{B}_n} s^{\operatorname{des}(w)} t^{\operatorname{des}(w^{-1})},$$

as studied by Visontai in [Vis13]. We clearly have

$$\mathbb{P}\big((D_n, D'_n) = (i,j)\big) = \frac{B_{n,i,j}}{|\mathtt{B}_n|}.$$

Lemma 5.19. *The numbers $B_{n,i,j}$ satisfy the following recursion formula:*

$$\begin{aligned} nB_{n,i,j} = &\left(n+i+j+2ij\right) B_{n-1,i,j} \\ &+ \left(1-i+\left(2n+1\right) j-2ij\right) B_{n-1,i-1,j} \\ &+ \left(1-j+\left(2n+1\right) i-2ij\right) B_{n-1,i,j-1} \\ &+ \left(n\left(2n+3\right)-\left(2n+1\right) i-\left(2n+1\right) j+2ij\right) B_{n-1,i-1,j-1}. \end{aligned} \tag{5.11.7}$$

Proof. In [Vis13, Theorem 15], Visontai shows that the type $\mathtt{B}_n$ two-sided Eulerian polynomial satisfies

$$\begin{aligned} nB_n\left(s,t\right) = &\left(2n^2st-nst+n\right) B_{n-1}\left(s,t\right) \\ &+ \left(2nst\left(1-s\right)+s\left(1-s\right)\left(1-t\right)\right) \frac{\partial}{\partial s} B_{n-1}\left(s,t\right) \\ &+ \left(2nst\left(1-t\right)+t\left(1-s\right)\left(1-t\right)\right) \frac{\partial}{\partial t} B_{n-1}\left(s,t\right) \\ &+ 2st\left(1-s\right)\left(1-t\right) \frac{\partial^2}{\partial s \partial t} B_{n-1}\left(s,t\right). \end{aligned}$$

From this, Eq. (5.11.7) follows by computing the derivatives and comparing the coefficients on both sides. □

Using this, we obtain the following analogue of Lemma 5.17:

Lemma 5.20. *In type $\mathtt{B}_n$, the random variable (D_n, D'_n) satisfies the following:*

$$\mathbb{E}\left(\left(D_{n+1}, D'_{n+1}\right) \mid \left(D_n, D'_n\right)\right) = \begin{cases} \left(D_n, D'_n\right) & \text{with prob.} \quad P_1 = \frac{n+1+D_n+D'_n+2D_nD'_n}{2(n+1)^2}, \\ \left(D_n+1, D'_n\right) & \text{with prob.} \quad P_2 = \frac{-D_n+(2n+1)D'_n-2D_nD'_n}{2(n+1)^2}, \\ \left(D_n, D'_n+1\right) & \text{with prob.} \quad P_3 = \frac{(2n+1)D_n-D'_n-2D_nD'_n}{2(n+1)^2}, \\ \left(D_n+1, D'_n+1\right) & \text{with prob.} \quad P_4 = \frac{(2n+1)(n+1-(D_n+D'_n))+2D_nD'_n}{2(n+1)^2}. \end{cases}$$

Proof. Dividing both sides of Eq. (5.11.7) by $n2^n n!$, we obtain

$$\begin{aligned} \frac{B_{n,i,j}}{\left|\mathtt{B}_n\right|} = &\frac{n+i+j+2ij}{2n^2} \frac{B_{n-1,i,j}}{\left|\mathtt{B}_{n-1}\right|} \\ &+ \frac{1-i+\left(2n+1\right) j-2ij}{2n^2} \frac{B_{n-1,i-1,j}}{\left|\mathtt{B}_{n-1}\right|} \\ &+ \frac{1-j+\left(2n+1\right) i-2ij}{2n^2} \frac{B_{n-1,i,j-1}}{\left|\mathtt{B}_{n-1}\right|} \\ &+ \frac{n\left(2n+3\right)-\left(2n+1\right) i-\left(2n+1\right) j+2ij}{2n^2} \frac{B_{n-1,i-1,j-1}}{\left|\mathtt{B}_{n-1}\right|}, \end{aligned}$$

where we used that $\left|\mathtt{B}_n\right| = 2^n n!$. From this, the result follows because, as noted above, we have

$$\frac{B_{n,i,j}}{\left|\mathtt{B}_n\right|} = \mathbb{P}\big(\left(D_n, D'_n\right) = \left(i,j\right)\big) \qquad \text{and} \qquad \frac{B_{n-1,k,l}}{\left|\mathtt{B}_{n-1}\right|} = \mathbb{P}\left(\left(D_{n-1}, D'_{n-1}\right) = \left(k,l\right)\right). \quad \square$$

Proposition 5.21. *In type* $\mathtt{B}_n$, $n \geq 4$, *the fourth centered moment of* T_n *is given by*

$$\mathbb{E}\left((T_n - \mathbb{E}(T_n))^4\right) = \frac{1}{60}\left(5n^2 + 39n + 79\right) + \frac{2n-1}{4n(n-1)}.$$

Proof. The proof is completely analogous to the one of Proposition 5.18. Again, set $U_n := D_n - \mathbb{E}(D_n)$ and $U_n' := D_n' - \mathbb{E}(D_n')$ such that $T_n - \mathbb{E}(T_n) = U_n + U_n'$ and observe that it suffices to compute $\mathbb{E}\left(U_n^k {U'}_n^l\right)$ for all $0 \leq k, l \leq 4$ with $k + l = 4$.

This can now inductively be done using the recursion formula

$$\mathbb{E}(U_{n+1}^k {U'}_{n+1}^{\ l}) = \mathbb{E}(U_n^k {U'}_n^{\ l} P_1 + (U_n+1)^k {U'}_n^{\ l} P_2 + U_n^k (U_n'+1)^l P_3 + (U_n+1)^k (U_n'+1)^l P_4),$$

where P_1, P_2, P_3 and P_4 are as in Lemma 5.20. We solved the corresponding recursions with the **RSolve** command of MATHEMATICA. Intermediate results can be found in Appendix A.2. □

5.11.4. Proof of Theorem 5.13

We are now able to formulate a proof for Theorem 5.13 by deriving the order of the fourth moment of $T_{\mathtt{D}_n}$ from the corresponding moment of $T_{\mathtt{B}_n}$ via the Minkowski inequality:

Proof of Theorem 5.13. For type $\mathtt{A}_n$ and $\mathtt{B}_n$, we obtained the result in Proposition 5.18 and Proposition 5.21, respectively. For type $\mathtt{D}_n$, we exploit the similarity of $\mathtt{B}_n$ and $\mathtt{D}_n$ to bound the difference between the respective fourth moments. The group $\mathtt{B}_n$ has a more combinatorial description as a group of signed permutations (for further details, see Section 5.2.2 or [BB05, Chapter 8]). Choosing an element of $\mathtt{B}_n$ uniformly at random hence is equivalent to choosing a random permutation $\pi \in \mathrm{Sym}(n)$ together with a tuple $(b_1, \ldots, b_n) \in \{\pm 1\}^n$. We then obtain $\tilde{\pi} \in \mathtt{B}_n$ by setting $\tilde{\pi}(i) := b_i \cdot \pi(i)$ as explained in Section 5.8. In this description, $\mathtt{D}_n$ is the subgroup of $\mathtt{B}_n$ given by all signed permutations $\tilde{\pi}$ such that $|\{i \in \{1, \ldots, n\} \mid \tilde{\pi}(i) < 0\}|$ is an even number. Choosing an element of $\tilde{\pi} \in \mathtt{D}_n$ uniformly at random is equivalent to choosing a random permutation $\pi \in \mathrm{Sym}(n)$ together with a tuple $(b_1, \ldots, b_{n-1}) \in \{\pm 1\}^{n-1}$ and setting

$$\tilde{\pi}(i) := \begin{cases} b_i \cdot \pi(i) & , 1 \leq i \leq n-1, \\ \left(\prod_{j=1}^{n-1} b_j\right) \cdot \pi(i) & , i = n. \end{cases}$$

These considerations imply that we can write

$$T_{\mathtt{D}_n} \stackrel{d}{=} T_{\mathtt{B}_n} + Y_n,$$

where Y_n is a bounded random variable (cf. proof of Theorem 5.12). Using the Minkowski inequality, we obtain

$$\begin{aligned} \mathbb{E}\left((T_{\mathtt{D}_n} - \mathbb{E}(T_{\mathtt{D}_n}))^4\right) &\leq \left(\left(\mathbb{E}\left(T_{\mathtt{B}_n} - \mathbb{E}(T_{\mathtt{B}_n})\right)^4\right)^{\frac{1}{4}} + O(1)\right)^4 \\ &= \mathbb{E}\left((T_{\mathtt{B}_n} - \mathbb{E}(T_{\mathtt{B}_n}))^4\right) + O\left(n^{\frac{3}{2}}\right). \end{aligned}$$

The result now follows from Proposition 5.21. □

Remark 5.22. The results of this section show the convenience of the conditional expectation to compute the expected value: Instead of a combinatorial approach as for example in the proof of [KS20, Proposition 5.7], one derives a recursion formula and uses a recursion solver like **RSolve** to find the solution. Of course, this approach is only possible if one can find a conditional expectation as for example in Lemma 5.20.

Remark 5.23. In [Ö19, Section 5.7] it is shown how to derive the CLT for T when $(W_n)_n = (\mathtt{A}_n)_n$ via the martingale convergence theorem and the recursive formulation of Lemma 5.17. This is an alternative proof of [CD17, Theorem 1.1] (see Theorem 5.10) and one should be able to find an alternative proof for Theorem 5.11, i.e to prove the CLT for T when $(W_n)_n = (\mathtt{B}_n)_n$ with the given formulas for the moments of $T_{\mathtt{B}}$.

5.12. CLTs for weighted sums of converging sequences

This section explains how to derive the asymptotic normality of a sequence of random variables $(X_n)_n$, where $X_n = \sum_{i=1}^{k_n} a_{n,i} X_{n,i}$, under the assumption that $(X_{n,i})_n \xrightarrow{D} N(0,1)$ for all i. The main idea is to use Lévy's continuity theorem via the pointwise convergence of the characteristic function of X_n towards the characteristic function of the standard normal distribution. We begin with some preparations:

Definition 5.24. The *characteristic function* of a random variable X is defined as $\psi_X(t) := \mathbb{E}\left(e^{itX}\right)$ for $t \in \mathbb{R}$.

For a detailed introduction to characteristic functions, see for example in [Bil95]. Now, Lévy's continuity theorem states the following:

Theorem 5.25 (Lévy)**.** *For a sequence of random variables $(X_n)_n$, it holds that $X_n \xrightarrow{D} X$ for some random variable X if and only if $\lim_{n\to\infty} \psi_{X_n}(t) = \psi_X(t)$ for every $t \in \mathbb{R}$.*

Characteristic functions of sums of independent random variables exhibit the following useful property:

Lemma 5.26. *Let X and Y be real-valued random variables. If X and Y are independent and $a, b \in \mathbb{R}$, it holds that $\psi_{aX+bY}(t) = \psi_X(at)\psi_Y(bt)$ for every $t \in \mathbb{R}$.*

Using the preceding results, one obtains the following lemma, which describes when a weighted sum of converging sequences satisfies the CLT.

Lemma 5.27. *Let $X_n = \sum_{i=1}^{k_n} a_{n,i} X_{n,i}$, where for every n, the $X_{n,i}$ are independent centered random variables with $\mathbb{V}(X_{n,i}) = 1$ and $a_{n,i} \in \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^{k_n} a_{n,i}^2 = 1$. Then if for each i, we have $X_{n,i} \xrightarrow{D} N(0,1)$ and*

$$\lim_{k\to\infty} \sup_n \left(\sum_{i=k}^{k_n} a_{n,i}^2 \right) = 0, \tag{5.12.1}$$

it follows that $X_n \xrightarrow{D} N(0,1)$.

Before we prove this statement, we give some comments on Eq. (5.12.1). Let $X_n^k := \sum_{i=1}^{\min(k,k_n)} a_{n,i}X_{n,i}$ be the random variable consisting of the first k summands of X_n. We have $\mathbb{V}(X_n) = \sum_{i=1}^{k_n} a_{n,i}^2 = 1$ and

$$\mathbb{V}(X_n^k) = \sum_{i=1}^{\min(k,k_n)} a_{n,i}^2 = 1 - \sum_{i=k}^{k_n} a_{n,i}^2.$$

Hence, Eq. (5.12.1) is equivalent to

$$\lim_{k\to\infty} \sup_n \left(\mathbb{V}(X_n) - \mathbb{V}(X_n^k)\right) = 0.$$

This means that the statement of Lemma 5.27 can roughly be phrased as follows: If all the columns of the array $(X_{n,i})_{n,i}$ satisfy the CLT and furthermore, the initial summands of X_n asymptotically contain all of the variance of X_n, then $(X_n)_n$ satisfies the CLT.

Proof of Lemma 5.27. The characteristic function of the standard normal distribution is $e^{-\frac{1}{2}t^2}$. To prove the asymptotic normality of X_n, we therefore show that for all $t \in \mathbb{R}$ and any $\delta > 0$, there is an $N \in \mathbb{N}$ so that $|\psi_{X_n}(t) - e^{-\frac{1}{2}t^2}| < \delta$ for all $n \geq N$. Now,

$$\begin{aligned}|\psi_{X_n}(t) - e^{-\frac{1}{2}t^2}| &= |\psi_{X_n}(t) - \psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t) + \psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t) - e^{-\frac{1}{2}t^2}| \\ &\leq |\psi_{X_n}(t) - \psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t)| + |\psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t) - e^{-\frac{1}{2}t^2}|.\end{aligned}$$

Eq. (5.12.1) guarantees that for any $\varepsilon > 0$, there is a finite k such that for all n, one has $\sum_{i=k+1}^{k_n} a_{n,i}^2 \leq \varepsilon$. We conclude for the first summand with Jensen's inequality and $|e^{i\alpha} - 1| \leq |\alpha|$, that

$$\begin{aligned}|\psi_{X_n}(t) - \psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t)| &= |\mathbb{E}(e^{itX_n} - e^{it\sum_{i=1}^k a_{n,i}X_{n,i}})| \\ &\leq \mathbb{E}|e^{it\sum_{i=k+1}^\infty a_{n,i}X_{n,i}} - 1| \\ &\leq \mathbb{E}|t\sum_{i=k+1}^{\infty} a_{n,i}X_{n,i}| \\ &\leq |t|\left(\mathbb{E}\left(\sum_{i=k+1}^{\infty} a_{n,i}X_{n,i}\right)^2\right)^{\frac{1}{2}} \leq |t|\left(\sum_{i=k+1}^{\infty} a_{n,i}^2\right)^{\frac{1}{2}} \leq |t|\varepsilon^{\frac{1}{2}}.\end{aligned}$$

For the second summand, with the uniform convergence of characteristic functions on compact intervals and the asymptotic normality of $(X_{n,i})_n$, i.e. $\psi_{X_{n,i}}(t) \to e^{-\frac{1}{2}t^2}$, we obtain for some positive constants C_1, C_2

$$\begin{aligned}|\psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t) - e^{-\frac{t^2}{2}}| &= |\psi_{\sum_{i=1}^k a_{n,i}X_{n,i}}(t) - e^{-\sum_{i=1}^k a_{n,i}^2\frac{t^2}{2}} + e^{-\sum_{i=1}^k a_{n,i}^2\frac{t^2}{2}} - e^{-\frac{t^2}{2}}| \\ &\leq |\prod_{i=1}^k \psi_{X_{n,i}}(a_{n,i}t) - \prod_{i=1}^k e^{-a_{n,i}^2\frac{t^2}{2}}| + |e^{-\sum_{i=1}^k a_{n,i}^2\frac{t^2}{2}} - e^{-\frac{t^2}{2}}|\end{aligned}$$

$$\leq C_1\varepsilon + |e^{-\frac{t^2}{2}}(e^{-(1-\sum_{i=1}^k a_{n,i}^2)\frac{t^2}{2}} - 1)|$$
$$\leq C_1\varepsilon + |e^{-\frac{t^2}{2}}(e^{-\varepsilon\frac{t^2}{2}} - 1)| \leq C_2\varepsilon.$$

These considerations imply that for any $\varepsilon > 0$ and some positive constant $C_3(t)$, there is an $N \in \mathbb{N}$ so that for all $n \geq N$ it holds that $|\psi_{X_n}(t) - e^{-\frac{1}{2}t^2}| \leq C_3(t)\varepsilon = \delta$. □

The following lemma is a consequence of Lemma 5.27 when k_n is globally bounded, but additionally allows for summands that converge in probability towards zero, instead of converging in distribution to the standard normal distribution.

Lemma 5.28. *Let $(X_n)_n$ be a sequence of centered random variables and suppose that there is $k \in \mathbb{N}$ such that for each n, X_n can be written as a sum $X_n = X_{n,1} + \cdots + X_{n,k}$ of independent random variables $X_{n,i}$. Assume that for every $1 \leq i \leq k$, the following holds true: Either $(X_{n,i})_n$ satisfies the CLT or $\frac{X_{n,i}}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{\mathbb{P}} 0$. Then if at least one sequence $(X_{n,i})_n$ satisfies the CLT and $\mathbb{V}(X_n) \to \infty$, the sequence $(X_n)_n$ satisfies the CLT.*

Proof. Without loss of generality, we can assume that there is a $k' \geq 1$ such that for $1 \leq i \leq k'$, the sequence $(X_{n,i})_n$ satisfies the CLT while for all $i > k'$, we have $\frac{X_{n,i}}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{\mathbb{P}} 0$. This implies that

$$Z_n := \frac{X_{n,k'+1} + \cdots + X_{n,k}}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{\mathbb{P}} 0$$

Using Slutsky's Theorem [Leh98, Theorem 2.3.3], we see that X_n satisfies the CLT if the remaining sum $X_n' = X_n - Z_n = X_{n,1} + \cdots + X_{n,k'}$ satisfies the CLT. We write

$$\frac{X_n'}{\sqrt{\mathbb{V}(X_n')}} = \sum_{i=1}^{k'} a_{n,i}\frac{X_{n,i}}{\sqrt{\mathbb{V}(X_{n,i})}}, \qquad \text{where} \qquad a_{n,i} = \sqrt{\frac{\mathbb{V}(X_{n,i})}{\mathbb{V}(X_n')}}.$$

Now, we have

$$\sum_{i=1}^{k'} a_{n,i}^2 = \frac{\sum_{i=1}^{k'}\mathbb{V}(X_{n,i})}{\mathbb{V}(X_n')} = 1,$$

so the claim follows from Lemma 5.27. Furthermore, Eq. (5.12.1) is trivially satisfied. □

Lemma 5.29. *In the setting of Lemma 5.28, $\frac{X_{n,i}}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{\mathbb{P}} 0$ holds if $\frac{\mathbb{V}(X_{n,i})}{\mathbb{V}(X_n)} \to 0$.*

Proof. The Chebyshev inequality shows that

$$\mathbb{P}\left(\frac{|X_{n,i}|}{\sqrt{\mathbb{V}(X_n)}} \geq \varepsilon\right) \leq \frac{\mathbb{V}(X_{n,i})}{\varepsilon^2\mathbb{V}(X_n)},$$

which implies the convergence in probability of $\frac{|X_{n,i}|}{\sqrt{\mathbb{V}(X_n)}}$ towards zero if $\frac{\mathbb{V}(X_{n,i})}{\mathbb{V}(X_n)} \to 0$. □

5.13. CLTs via the Lindeberg Theorem

A collection $(X_{n,i})_{n\geq 1}^{1\leq i\leq k_n}$ of random variables is a *triangular array* if for each n, all $X_{n,i}$ are independent of each other. A triangular array is *centered* if $\mathbb{E}(X_{n,i}) = 0$ for all n and i. Given such a triangular array, we set

$$X_n \coloneqq \sum_{i=1}^{k_n} X_{n,i}, \qquad s_{n,i}^2 \coloneqq \mathbb{V}(X_{n,i}) \qquad \text{and} \qquad s_n^2 \coloneqq \mathbb{V}(X_n) = \sum_{i=1}^{k_n} s_{n,i}^2.$$

The array $(X_{n,i})_{n,i}$ satisfies the *maximum condition* if

$$\lim_{n\to\infty} \max_{1\leq i\leq k_n} \frac{s_{n,i}^2}{s_n^2} = 0.$$

It satisfies the *Lindeberg condition* if for every $\varepsilon > 0$,

$$\frac{1}{s_n^2} \sum_{i=1}^{k_n} \mathbb{E}\left(X_{n,i}^2 \mathbb{1}_{\{|X_{n,i}|>\varepsilon s_n\}}\right) \to 0.$$

The importance of these conditions for us is as follows:

Theorem 5.30 (Lindeberg)**.** *Let $(X_{n,i})_{n,i}$ be a centered triangular array. Then $(X_{n,i})_{n,i}$ satisfies the Lindeberg condition if and only if it satisfies the maximum condition and the sequence $(X_n)_n$ satisfies the CLT.*

To apply this to our setting, let $(W_n)_n$ be a sequence of finite Coxeter groups and let

$$W_n = \prod_{i=1}^{k_n} W_{n,i},$$

be the decomposition of W_n into its irreducible components. Now, let T_n be the random variable associated to the statistic t on W_n. By Lemma 5.4, we have

$$T_n = \sum_{i=1}^{k_n} T_{n,i},$$

where $T_{n,i}$ is the random variable associated to the statistic t on $W_{n,i}$. From this, we obtain a centered triangular array by setting $X_{n,i} \coloneqq T_{n,i} - \mathbb{E}(T_{n,i})$. By the arguments above, we have $X_n = T_n - \mathbb{E}(T_n)$.

As a first application of the Lindeberg Theorem, we obtain a CLT for products of dihedral groups:

Lemma 5.31. *Let $(W_n)_n$ be a sequence of finite Coxeter groups such that for each n, every irreducible component of W_n is of dihedral type. Write*

$$W_n = \prod_{i=1}^{k_n} \mathtt{I}_2(m_{n,i})$$

and let T_n be the random variable associated to the statistic t on W_n. Then if $\sum_{i=1}^{k_n} \frac{1}{m_{n,i}} \to \infty$, the sequence $(T_n)_n$ satisfies the CLT.

Proof. Define the triangular array $(X_{n,i})_{n,i}$ associated to the sequence $(W_n)_n$ as explained above. We want to show that this array satisfies both the maximum condition and the Lindeberg condition.

By Theorem 5.5, we have for all n and i

$$s_{n,i}^2 = \mathbb{V}(X_{n,i}) = 4/m_{n,i} \leq 4/3$$

and $s_n^2 = \sum_{i=1}^{k_n} 4/m_{n,i}$. Thus, it follows immediately from the assumption that the maximum condition is satisfied.

It is easy to verify that for all n and i and all $w \in \mathtt{I}_2(m_{n,i})$, one has

$$0 \leq t(w) = \mathrm{des}(w) + \mathrm{des}(w^{-1}) \leq 4$$

as this is true for all dihedral groups. We have $\mathrm{rk}(\mathtt{I}_2(m_{n,i})) = 2$, so by Theorem 5.5, one has

$$|X_{n,i}| = |T_{n,i} - \mathbb{E}(T_{n,i})| \leq 2.$$

By assumption, $s_n \to \infty$, so for every $\varepsilon > 0$, the indicator function $\mathbb{1}_{\{|X_{n,i}| > \varepsilon s_n\}}$ is trivial for n sufficiently large. This implies that $X_{n,i}$ satisfies the Lindeberg condition.

Now by Theorem 5.30, the sequence $(X_n)_n = (T_n - \mathbb{E}(T_n))_n$ satisfies the CLT and hence so does $(T_n)_n$. □

We obtain the following result for sequences of Coxeter groups with no dihedral irreducible components:

Lemma 5.32. *Let $(W_n)_n$ be a sequence of finite Coxeter groups such that for each n, every irreducible component $W_{n,i}$ of W_n is of non-dihedral type and we have $\mathrm{rk}(W_{n,1}) \geq \ldots \geq \mathrm{rk}(W_{n,k_n})$. Then if $\mathrm{rk}(W_{n,1}) = o(\mathrm{rk}(W_n))$, the random variable T_n associated to the statistic t on W_n satisfies the CLT.*

Proof. As above, let $(X_{n,i})_{n,i}$ be the triangular array associated to the sequence $(W_n)_n$. By Theorem 5.5, we know that s_n^2 is of the order of $\mathrm{rk}(W_n)$ and $s_{n,i}^2$ is of order $\mathrm{rk}(W_{n,i})$. Therefore, the maximum condition is satisfied, as $\max_{1 \leq i \leq k_n} \mathrm{rk}(W_{n,i}) = \mathrm{rk}(W_{n,1}) = o(\mathrm{rk}(W_n))$. With the Cauchy–Schwarz inequality, the Chebyshev inequality and the results for the fourth moment from Theorem 5.13, we see that

$$\begin{aligned}\mathbb{E}\left(\frac{X_{n,i}^2}{s_{n,i}^2}\mathbb{1}_{\{|X_{n,i}| > \varepsilon s_n\}}\right) &\leq \sqrt{\frac{\mathbb{E}(X_{n,i}^4)}{s_{n,i}^4}\mathbb{P}(|X_{n,i}| > \varepsilon s_n)} \\ &= O\left(\frac{s_{n,i}}{s_n}\right).\end{aligned}$$

The factors $W_{n,i}$ which are of exceptional type can be neglected here since for them, $|X_{n,i}|$ is globally bounded. This implies

$$\frac{1}{s_n^2}\sum_{i=1}^{k_n} \mathbb{E}\left(X_{n,i}^2 \mathbb{1}_{\{|X_{n,i}| > \varepsilon s_n\}}\right) = O\left(\frac{1}{s_n^2}\sum_{i=1}^{k_n} s_{n,i}^2 \frac{s_{n,i}}{s_n}\right)$$

$$= O\left(\frac{s_{n,1}}{s_n}\sum_{i=1}^{k_n}\left(\frac{s_{n,i}}{s_n}\right)^2\right)$$
$$= O\left(\frac{s_{n,1}}{s_n}\left(\sum_{i=1}^{k_n}\frac{s_{n,i}}{s_n}\right)^2\right) = O\left(\frac{s_{n,1}}{s_n}\right),$$

where we assumed that, without loss of generality, for each n, we have $s_{n,1} = \sqrt{\mathbb{V}(T_{n,1})} \geq s_{n,i}$ for all i. The CLT now follows because, as observed above, we have $s_{n,1} = o(s_n)$. □

5.14. Proof of the main theorem

Throughout this section, let $(W_n)_n$ be a sequence of finite Coxeter groups such that $\mathrm{rk}(W_n) \to \infty$, let

$$W_n = \prod_{i=1}^{k_n} W_{n,i}$$

be the decomposition of W_n into its irreducible components and assume that for all n, we have $\mathrm{rk}(W_{n,1}) \geq \ldots \geq \mathrm{rk}(W_{n,k_n})$. As above, let $T_n := T_{W_n}$ and $T_{n,i} := T_{W_{n,i}}$.

In the previous section, we proved the CLT for sequences where either every $W_{n,i}$ is of dihedral type (Lemma 5.31) or where every $W_{n,i}$ is of non-dihedral type and $\mathrm{rk}(W_{n,i}) = o(\mathrm{rk}(W_n))$ (Lemma 5.32). The proofs required a maximum condition: We used that in both cases, the variance of $T_{n,i}$ was of smaller magnitude than the variance of T_n. However, this need not be the case in general. If the $W_{n,i}$ are of non-dihedral type, it is possible that for some i, the rank of $W_{n,i}$ is of the same order as the rank of W_n. An easy example of this is given by setting $W_n := \mathtt{A}_n^k$ for some $k \in \mathbb{N}$. Here, we have $\mathbb{V}(T_n)/\mathbb{V}(T_{n,i}) = k$ for all n. An example with a growing number of irreducible components is the sequence $W_n = \prod_{i=1}^{\lceil \log(n) \rceil} A_{\lceil \frac{n}{2^i} \rceil}$, so that $\mathbb{V}(T_n)/\mathbb{V}(T_{n,i}) = 2^i$. In order to extend our results to these cases, we need to separate the irreducible components that do not satisfy the maximum condition from the remaining ones. For this, we make the following definition:

An irreducible component $W_{n,i}$ of W_n is *δ-small* for some $\delta > 0$, if $\mathrm{rk}(W_{n,i}) \leq \mathrm{rk}(W_n)^{1-\delta}$. Let $m_n := \min\{i \in \mathbb{N} : \ W_{n,i+1} \text{ is } \delta\text{-small}\}$. Define $M_n^\delta := \prod_{i=1}^{m_n} W_{n,i}$ and $W_n^\delta := \prod_{i=m_n+1}^{k_n} W_{n,i}$. For all n, we can write $W_n = M_n^\delta \times W_n^\delta$. By Lemma 5.4, we have

$$T_n = T_{M_n^\delta} + T_{W_n^\delta} = \sum_{i=1}^{m_n} T_{n,i} + \sum_{i=m_n+1}^{k_n} T_{n,i}.$$

We note that if every $W_{n,i}$ is of non-dihedral type and for some δ, one has $\lim_{n\to\infty} m_n = 0$, the maximum condition is satisfied.

Remark 5.33. Every dihedral group has rank 2 and every finite irreducible Coxeter group of exceptional type has rank smaller than 9. Hence, for every $0 < \delta < 1$, there

is $N \in \mathbb{N}$ such that for all $n \geq N$, every irreducible components of W_n is either of type $\mathtt{A}, \mathtt{B}$ or $\mathtt{D}$ or it is δ-small.

As was shown by Chatterjee–Diaconis [CD17] (see Theorem 5.10) and Röttger [Röt20] (see Theorem 5.11 and Theorem 5.12), the sequences $T_{\mathtt{A}_n}, T_{\mathtt{B}_n}$ and $T_{\mathtt{D}_n}$ satisfy the CLT. This allows us to apply Lemma 5.27 if the sequence $(W_n)_n$ satisfies the following property:

Definition 5.34. Let $W_n = M_n^\delta \times W_n^\delta$ as defined above. The sequence $(W_n)_n$ is *well-behaved*, if there exists some $\delta > 0$, so that

$$\lim_{k\to\infty} \sup_n \left(\sum_{i=k}^{m_n} \frac{\mathbb{V}(T_{n,i})}{\mathbb{V}(T_{M_n^\delta})} \right) = 0. \tag{5.14.1}$$

While the definition seems to be rather technical, we have failed to construct a sequence of finite Coxeter groups that is not well-behaved. Some examples to illustrate this are listed in Example 5.38.

Remark 5.35. For all $J \subseteq \mathbb{N}$, we obviously have

$$\sup_{n\in J} \left(\sum_{i=k}^{m_n} \frac{\mathbb{V}(T_{n,i})}{\mathbb{V}(T_{M_n^\delta})} \right) \leq \sup_{n\in\mathbb{N}} \left(\sum_{i=k}^{m_n} \frac{\mathbb{V}(T_{n,i})}{\mathbb{V}(T_{M_n^\delta})} \right) \text{ for all } k.$$

Thus, every subsequence of a well-behaved sequence is well-behaved again.

Proposition 5.36. *If $(W_n)_n$ is well-behaved, and all $W_{n,i}$ are of non-dihedral type, then the sequence $(T_n)_n$ satisfies the CLT.*

Proof. Choose δ such that Eq. (5.14.1) is satisfied. As noted above, we have $T_n = T_{M_n^\delta} + T_{W_n^\delta}$. From Theorem 5.5, we know that $\mathbb{V}(T_{M_n^\delta})$ is of order $\mathrm{rk}(M_n^\delta)$ and $\mathbb{V}(T_{W_n^\delta})$ is of order $\mathrm{rk}(W_n^\delta)$. By assumption, we have $\mathrm{rk}(W_n) = \mathrm{rk}(M_n^\delta) + \mathrm{rk}(W_n^\delta) \to \infty$.

By Lemma 5.1, it suffices to show that every subsequence of $(T_n)_n$ has a subsequence which satisfies the CLT. For any $J \subseteq \mathbb{N}$, the subsequence $(W_n)_{n\in J}$ satisfies all conditions of the proposition: $(W_n)_{n\in J}$ is a sequence of finite Coxeter groups which have no irreducible factors of dihedral type and such that $(\mathrm{rk}(W_n))_{n\in J}$ tends to infinity. Furthermore, this subsequence is well-behaved as noted in Remark 5.35. Thus, we can assume that $J = \mathbb{N}$, i.e. it suffices to show that $(T_n)_{n\in\mathbb{N}}$ has a subsequence which satisfies the CLT.

If $\mathrm{rk}(M_n^\delta) = o(\mathrm{rk}(W_n))$ then $\mathrm{rk}(W_n^\delta)$ must be of the same order as $\mathrm{rk}(W_n)$. Hence, as every irreducible factor of $W_n^\delta = \prod_{i=m_n+1}^{k_n} W_{n,i}$ is δ-small, we have $\mathrm{rk}(W_{n,m_{n+1}}) = o(\mathrm{rk}(W_n^\delta))$. This allows us to apply Lemma 5.32 to see that $(T_{W_n^\delta})_n$ satisfies the CLT. The CLT for $(T_n)_n$ now follows, even without passing to a subsequence, from Lemma 5.28 and Lemma 5.29 because $\mathbb{V}(T_{M_n^\delta})/\mathbb{V}(T_n) \to 0$.

Next assume that $\mathrm{rk}(M_n^\delta) \neq o(\mathrm{rk}(W_n))$. In this case, there is $J \subseteq \mathbb{N}$ such that $(\mathrm{rk}(M_n^\delta))_{n\in J} \to \infty$. The subsequence $(M_n^\delta)_{n\in J}$ is again well-behaved and as noted in Remark 5.33, we can assume that every irreducible component of M_n^δ is of type $\mathtt{A}, \mathtt{B}$ or $\mathtt{D}$. Thus, it follows from [CD17],[Röt20] and Lemma 5.27 that the sequence $(T_{M_n^\delta})_{n\in J}$

satisfies the CLT. There are two cases to consider: If $\mathrm{rk}(W_n^\delta) = o(\mathrm{rk}(W_n))$, we have $\mathbb{V}(T_{W_n^\delta})/\mathbb{V}(T_n) \to 0$. If this is not the case, it follows from Lemma 5.32 that, after possible passing to a further subsequence, $T_{W_n^\delta}$ satisfies the CLT. In both cases, the asymptotic normality of $(T_n)_{n\in J}$ follows from Lemma 5.28 and Lemma 5.29. □

We are now ready to prove our main theorem. Each W_n decomposes uniquely as

$$W_n = G_n \times I_n,$$

where no irreducible component of G_n is of dihedral type and

$$I_n = \prod_{i=1}^{l_n} \mathtt{I}_2(m_{n,i}).$$

Note that by Remark 5.33, the sequence $(W_n)_n$ is well-behaved if and only if $(G_n)_n$ is. We use this decomposition in order to combine the results obtained so far and show:

Theorem 5.37. *Let T_n be the random variable associated to the statistic t on W_n. Assume that $(W_n)_n$ is well-behaved. Then the following are equivalent:*

1. *$(T_n)_n$ satisfies the CLT.*
2. *$\mathbb{V}(T_n) \to \infty$.*
3. *$\mathrm{rk}(G_n) + \sum_{i=1}^{l_n} \frac{1}{m_{n,i}} \to \infty$.*

Proof. By Lemma 5.4, the random variable T_n decomposes as a sum of independent random variables $T_n = T_n^G + T_n^I$, where $T_n^G = T_{G_n}$ and $T_n^I = T_{I_n}$. Let $r_n := \mathrm{rk}(G_n)$ and $d_n := \sum_{i=1}^{l_n} \frac{1}{m_{n,i}}$. By Theorem 5.5, r_n is of order $\mathbb{V}(T_n^G)$ and d_n is of order $\mathbb{V}(T_n^I)$. Using additivity of the variance, it follows immediately that Item 2 is equivalent to Item 3.

Now assume that Item 2 and Item 3 hold. We want to show that this implies Item 1. By Lemma 5.1, it suffices to show that every subsequence of $(T_n)_n$ has a subsequence which satisfies the CLT. For any $J \subseteq \mathbb{N}$, the subsequence $(W_n)_{n\in J}$ satisfies all conditions of the theorem and Item 2: $(W_n)_{n\in J}$ is a sequence of finite Coxeter groups such that both $(\mathrm{rk}(W_n))_{n\in J}$ and $(\mathbb{V}(T_n))_{n\in J}$ tend to infinity. Furthermore, this subsequence is well-behaved as noted in Remark 5.35. Thus, we can assume that $J = \mathbb{N}$ and have to show that $(T_n)_{n\in\mathbb{N}}$ has a subsequence which satisfies the CLT. If neither r_n nor d_n are bounded, there is $J \subseteq \mathbb{N}$ such that $(r_n)_{n\in J} \to \infty$ and $(d_n)_{n\in J} \to \infty$. By Proposition 5.36, $r_n \to \infty$ implies that T_n^G satisfies the CLT and by Lemma 5.31, $d_n \to \infty$ implies that T_n^I satisfies the CLT. Hence in this case, both $(T_n^G)_{n\in J}$ and $(T_n^I)_{n\in J}$ satisfy the CLT, so the subsequence $(T_n)_{n\in J}$ satisfies the CLT by Lemma 5.28. If r_n is bounded, d_n must be unbounded. Thus, we can find $J \subseteq \mathbb{N}$ such that $(d_n)_{n\in J} \to \infty$. It follows that $(T_n^I)_{n\in J}$ satisfies the CLT and that

$$\left(\frac{\mathbb{V}(T_n^G)}{\mathbb{V}(T_n)}\right)_{n\in J} \to 0,$$

so we can use Lemma 5.28 and Lemma 5.29 to see that $(T_n)_{n\in J}$ satisfies the CLT. The case where d_n is bounded works the same.

Lastly, as $T_n - \mathbb{E}(T_n)$ takes only values in $\mathbb{Z}$, the sequence $(T_n)_n$ can only satisfy a CLT if its variance tends to infinity [KS20, Proposition 6.15]. This shows that Item 1 implies Item 2. □

Example 5.38. The following list of examples illustrates Theorem 5.37. To simplify the notation, we omit the rounding of the ranks of the irreducible components.

- $W_n = \prod_{i=1}^{\log(n)} \mathtt{A}_{\frac{n}{2^i}} \times (\mathtt{B}_{\sqrt{n}})^{\sqrt{n}}$ satisfies the CLT, as $\prod_{i=1}^{\log(n)} \mathtt{A}_{\frac{n}{2^i}}$ is well-behaved, as $\frac{\mathbb{V}(T_{W_{n,i}})}{\mathbb{V}(T_{M_n^\delta})}$ does not depend on n, and $(\mathtt{B}_{\sqrt{n}})^{\sqrt{n}}$ satisfies the maximum condition.
- $\mathtt{B}_n \times (\mathtt{A}_{n^{1-\delta}})^{n^\delta}$ for any $0 < \delta < 1$ satisfies the CLT, as $m_n = 1$ is bounded and $(\mathtt{A}_{n^{1-\delta}})^{n^\delta}$ satisfies the maximum condition.
- $W_n = \prod_{i=1}^{n} I_2(i)$ satisfies the CLT, as the harmonic series diverges.
- $W_n = \prod_{i=1}^{n} I_2(i^2)$ does not satisfy the CLT.
- $W_n = \mathtt{A}_3^n \times \mathtt{D}_5^n \times \mathtt{F}_4^n \times I_2(n^2)$ satisfies the CLT.

Remark 5.39. After posting Theorem 5.37 to the arXiv (see [BR19]), V. Féray announced to us that the condition of the sequence of finite Coxeter groups to be well-behaved can indeed be dropped [Fé20]. For this, he employs a result of Mallows [Mal72] that allows him to use the L^2 Wasserstein distance to derive the CLT for those irreducible components that are of large variance. This is similar to our approach where we separate the irreducible components that are not δ-small. Féray's approach leads to a shorter proof of Theorem 5.37 without requiring the well-behaved condition.

5.15. Further results and outlook

There are different directions to extend the results that are presented in this chapter. For a start, it still remains to show that every sequence of finite Coxeter groups of growing rank is well-behaved or alternatively find a sequence that is not. A general question is to obtain a similar result for other statistics as we found for the two-sided descent statistic. Following Chatterjee and Diaconis [CD17], we show in Section 5.15.1 that the results of Section 5.9 and Section 5.10 extend to a class of statistics with bounded local degree. Furthermore, we show in Section 5.15.2 that a two-dimensional CLT for the statistic (D_n, D'_n) follows with the Theorem of Cramér–Wold for the irreducible types $\mathtt{B}_n$ and $\mathtt{D}_n$. A follow-up problem is the extension of Theorem 5.37 to a wider class of statistics, or, to specify when it may be extended and where our methods do not apply. A starting point is the peaks statistic, which should show a similar behavior as descents. A direct reproduction of Theorem 5.37 would require similar results on the fourth moments of other statistics as we found them for the two-sided descent statistic. If this cannot be done recursively, these will require new methods and ideas. As the more general

proof found by Féray [Fé20] only requires the variance instead of the fourth moments, his proof should apply to all two-sided statistics that satisfy the structure described in Section 5.15.1. An open problem of a different flavor suggested by Alperen Özdemir are Berry-Esseen bounds for the two-sided descent statistic. An interesting question is if it is possible to reproduce the bound for Wasserstein distance that hold for the irreducible types to arbitrary sequences of Coxeter groups. We were asked by Kyle Petersen if there is a two-dimensional CLT for the statistic (D_n, D'_n) for arbitrary sequences of Coxeter groups. We believe that this should follow from a careful application of the Cramér–Wold device to the proof of Theorem 5.37.

5.15.1. Generalization to a class of statistics with local degree k

As in [CD17], it is possible to generalize the proof of Theorem 5.11 to a wider class of statistics of local degree k. These statistics are of the form

$$F_1(\pi) + F_2(\pi^{-1}),$$

where the local components' absolute value is bounded by 1. If π is a signed permutation, a bound for the Wasserstein distance between the normalized statistic and the standard normal distribution follows. Therefore the central limit theorem for these statistics holds, if the variance of the statistics is of order $O(n^{\frac{1}{2}+\varepsilon})$ for an $\varepsilon > 0$. The Theorem is implied from a generalization of the proof of Theorem 5.11 by constructing the symmetric interaction rule in the right way.

Theorem 5.40. *Let W_n be a sequence of growing rank of Coxeter groups of type* B *and let F_1, F_2 be statistics of local degree k, with the absolute value of their local components bounded by 1. The statistic $F_1(\pi) + F_2(\pi^{-1})$ gives rise to a random variable F. The Wasserstein distance between F, normalized by its mean and variance, and the standard normal distribution satisfies*

$$\delta_F \leq C(k) \left(\frac{\sqrt{n}}{s^2} + \frac{n}{s^3} \right)$$

for $s^2 := \mathbb{V}(F_1(\pi) + F_2(\pi^{-1}))$ and some constant $C(k)$.

Proof. If the statistics F_1 and F_2 are of local degree k and their local components' absolute value is bounded by 1, let $\{i, j\}$ be an edge in $G(x)$ if and only if the x-ranks or the y-ranks differ by at most $k - 1$. For the extension $G'(x)$, we say that $\{i, j\}$ is an edge if and only if the ranks differ by at most $k + 3$. Then, Theorem 5.9 applies, and the Wasserstein distance is bounded:

$$\delta_F \leq C(k) \left(\frac{\sqrt{n}}{s^2} + \frac{n}{s^3} \right).$$

Here, $C(k)$ is a large enough constant. □

To see that the bound in Theorem 5.40 also holds when π is an element of a Coxeter group of type $\mathtt{D}_n$, we use the same technique as in the proof of Theorem 5.12. Hence, we decompose the statistic into a part that is the same statistic depending on a signed permutation on $\{\pm 1, \dots, \pm(n-1)\}$ and a finitely bounded random variable.

Theorem 5.41. *Let W_n be a sequence of growing rank of Coxeter groups of type $\mathtt{D}$ and let F_1, F_2 be statistics of local degree k, with the absolute value of their local components bounded by 1. The statistic $F_1(\pi) + F_2(\pi^{-1})$ gives rise to a random variable F. Then, if we assume that $\mathbb{V}(F) \to \infty$, the Wasserstein distance between F, normalized by its mean and variance, and the standard normal distribution satisfies*

$$\delta_F \leq C(k) \left(\frac{\sqrt{n-1}}{s^2} + \frac{n-1}{s^3} \right) + o(1)$$

for $s^2 := \mathbb{V}(F_1(\pi) + F_2(\pi^{-1}))$ and some constant $C(k)$.

Proof. Let $F = F_1(\pi_1) + F_2(\pi_1^{-1}) = f(X)$ where π_1 is a uniformly chosen element of the Coxeter group of type $\mathtt{D}_n$. Let $X = (X_1, \dots, X_n)$ be generated as in the proof of Theorem 5.12, so $X_i = (U_i, V_i, B_i)$ with $(U_i, V_i) \sim \mathrm{Unif}([0,1]^2)$. B_i is an independent random sign for $1 \leq i \leq n-1$ and $B_n = \prod_{i=1}^{n-1} B_i$. Then, F' is the statistic where we remove all local components that depend on B_n. Subsequently we add local components, so that the resulting statistic is $F' = F_1(\pi_2) + F_2(\pi_2^{-1})$, where π_2 is a random signed permutation on $\{\pm 1, \dots \pm (n-1)\}$ generated by $(X_1, \dots, X_{n-1})$. Then, as the local degree is k, $F - F' = O(1)$ and therefore $\mathbb{E}(F - F') = O(1)$ and $\mathbb{V}(F - F') = O(1)$, which implies that $\mathbb{V}(F') = \mathbb{V}(F) + O(1)$. Now, see that Eq. (5.10.1) from the proof of Theorem 5.12 generalizes to

$$\frac{F - \mathbb{E}(F)}{\sqrt{\mathbb{V}(F)}} = \frac{\sqrt{\mathbb{V}(F')}}{\sqrt{\mathbb{V}(F)}} \frac{F' - \mathbb{E}(F')}{\sqrt{\mathbb{V}(F')}} + \frac{F - F' - \mathbb{E}(F - F')}{\sqrt{\mathbb{V}(F)}},$$

which immediately shows that the Wasserstein distance between F and F' tends to zero, as $\lim_{n\to\infty} \frac{\mathbb{V}(F')}{\mathbb{V}(F)} = 1$ and $\lim_{n\to\infty} \frac{F - F' - \mathbb{E}(F - F')}{\sqrt{\mathbb{V}(F)}} = 0$. Therefore it holds that $\delta_F \leq \delta_{F'} + o(1)$ and the theorem follows. □

5.15.2. The Statistic $(\mathrm{des}(\pi), \mathrm{des}(\pi^{-1}))$

This section derives a two-dimensional central limit theorem for the vector statistic $(\mathrm{des}(\pi), \mathrm{des}(\pi^{-1}))$ for π being either an element of a Coxeter group of type $\mathtt{B}_n$ or $\mathtt{D}_n$. This is achieved with the Cramér–Wold device and a slight modification of the proofs of Theorems 5.11 and 5.12. The Cramér–Wold device shows the equivalence of the convergence in distribution between a random vector and every linear combination of its elements. It is also known as the Theorem of Cramér–Wold (see for example in [Bil95, Theorem 29.4]).

Theorem 5.42 (Cramér–Wold)**.** *Let $\bar{X}_n = (X_{n1}, \dots, X_{nk})$ and $\bar{X} = (X_1, \dots, X_k)$ be random vectors of dimension k. Then, $\bar{X}_n \xrightarrow{D} \bar{X}$, if and only if*

$$\sum_{i=1}^{k} t_i X_{ni} \xrightarrow{D} \sum_{i=1}^{k} t_i X_i$$

for each $t = (t_1, \dots, t_k) \in \mathbb{R}^k$ and for $n \to \infty$.

We use the short-hand notation (D_n, D'_n) for the random variable that rises from $(\mathrm{des}(\pi), \mathrm{des}(\pi^{-1}))$ as explained in Section 5.3. Therefore, we can show the convergence of (D_n, D'_n) by studying linear combinations of the form $t_1 D_n + t_2 D'_n$. It is sufficient to only check linear combinations with $t \in S^1$, since the investigated statistic is normalized by the square root of the variance $\mathbb{V}(t_1 D_n + t_2 D'_n)$. This leads to the following theorem:

Theorem 5.43. *Let W_n be a sequence of Coxeter groups of growing rank of either type $\mathtt{B}_n$ or $\mathtt{D}_n$. Then, the statistic (D_n, D'_n) satisfies a two-dimensional central limit theorem of the form*

$$\Sigma_n^{-\frac{1}{2}} \begin{pmatrix} D_n - \mathbb{E}(D_n) \\ D'_n - \mathbb{E}(D'_n) \end{pmatrix} \xrightarrow{D} N_2(0, I)$$

for $n \to \infty$, where I denotes the two-dimensional identity matrix and Σ_n is the covariance matrix of (D_n, D'_n).

Proof. Via the Theorem of Cramér–Wold, we can study the convergence of (D_n, D'_n) by studying $t_1 D_n + t_2 D'_n$ for $t^T = (t_1, t_2) \in S^1$. We derive a convergence

$$t^T \frac{1}{\sqrt{\mathbb{V}(D_n)}} \begin{pmatrix} D_n - \mathbb{E}(D_n) \\ D'_n - \mathbb{E}(D'_n) \end{pmatrix} \xrightarrow{D} N(0, 1) \tag{5.15.1}$$

to show the Theorem via an application of Slutsky's Theorem. (5.15.1) is equivalent to

$$\frac{1}{\sqrt{\mathbb{V}(D_n)}} (t_1 D_n + t_2 D'_n - (t_1 + t_2)\mathbb{E}(D_n)) \xrightarrow{D} N(0, 1), \tag{5.15.2}$$

as $\mathbb{E}(D_n) = \mathbb{E}(D'_n)$. Now, since $t \in S^1$, the proofs of the Theorems 5.11 and 5.12 apply, which means that

$$\frac{t_1 D_n + t_2 D'_n - (t_1 + t_2)\mathbb{E}(D_n)}{\sqrt{\mathbb{V}(t_1 D_n + t_2 D'_n)}} \xrightarrow{D} N(0, 1).$$

This convergence is also a consequence of Theorem 5.40 or Theorem 5.41, as the local components of $t_1 D_n + t_2 D'_n$ are still bound by 1 and the local dependency structure is not changed by multiplying the sum of indicator functions that model D_n and D'_n with constants. Furthermore, the variance $\mathbb{V}(t_1 D_n + t_2 D'_n)$ is of order n and therefore, the Wasserstein distance to the standard normal distribution is bound by a vanishing function in n. Now, by Slutsky's Theorem, (5.15.2) and therefore (5.15.1) is satisfied as

$$\frac{\mathbb{V}(t_1 D_n + t_2 D'_n)}{\mathbb{V}(D_n)} \xrightarrow{a.s} 1.$$

This results from the fact that $\mathbb{V}(D_n) = \mathbb{V}(D'_n))$ and $\mathrm{Cov}(D_n, D'_n) = O(1)$ (see [KS20]) and that $t_1^2 + t_2^2 = 1$. Because of the convergence in (5.15.1), the theorem follows via another application of Slutsky's Theorem, as

$$\begin{aligned}\frac{1}{\mathbb{V}(D_n)}\Sigma_n &= \frac{1}{\mathbb{V}(D_n)}\begin{pmatrix}\mathbb{V}(D_n) & \mathrm{Cov}(D_n, D'_n)\\ \mathrm{Cov}(D_n, D'_n) & \mathbb{V}(D'_n)\end{pmatrix}\\ &\xrightarrow{a.s.} I,\end{aligned}$$

since $\mathrm{Cov}(D_n, D'_n) = O(1)$ and $\mathbb{V}(D_n) = \mathbb{V}(D'_n)$. □

Remark 5.44. Theorem 5.43 can be generalized to certain statistics $(F_1(\pi), F_2(\pi^{-1}))$, if F_1 and F_2 meet the constraints of Theorem 5.40 or Theorem 5.41, $\mathbb{V}(F_1(\pi)) = \mathbb{V}(F_2(\pi^{-1}))$ holds and $\mathbb{V}(F_1(\pi))$ is big enough so that the constraint to the Wasserstein distance in Theorem 5.40 or Theorem 5.41 converges to zero for n going to infinity.

A. Moments of T

This section contains the moments up to degree 4 of the random variables which where described in the proofs of Proposition 5.18 and Proposition 5.21.

Let $D_n = D_{W_n}$, $T_n = T_{W_n}$, let D'_n be the random variable associated to the statistic

$$\begin{aligned} W_n &\to \mathbb{N} \\ w &\mapsto \mathrm{des}(w^{-1}) \end{aligned}$$

and define $U_n := D_n - \mathbb{E}(D_n)$ and $U'_n := D'_n - \mathbb{E}(D'_n)$. For the proofs of Proposition 5.18 and Proposition 5.21, one needs to inductively compute $\mathbb{E}(U_n^k {U'_n}^l)$ for all $0 \leq k, l \leq 4$ where $W_n = \mathtt{A}_n$ and $W_n = \mathtt{B}_n$, respectively. Note that $\mathbb{E}(U_n^k {U'_n}^l) = \mathbb{E}(U_n^l {U'_n}^k)$. For the sake of completeness, we also list the mixed moments of (D_n, D'_n), which can be computed similarly, although they are not needed to prove Proposition 5.18 and Proposition 5.21.

A.1. Type A

If $W_n = \mathtt{A}_n$, we obtain the following list of (joint) moments up to degree 4. The result for $\mathbb{E}(U_n^4)$ corresponds to Lemma 5.15 and the result for $\mathbb{E}((T_n - \mathbb{E}(T_n))^4)$ to Proposition 5.18. The moments in boldface were already known before and can be found in [KS20].

	$\mathbb{E}(\cdot)$
$\mathbf{U_n}$	0
$\mathbf{U_n^2}$	$\frac{n+2}{12}$
$\mathbf{U_n U'_n}$	$\frac{n}{2(n+1)}$
U_n^3	0
$U_n^2 U'_n$	0
$U_n^3 U'_n$	$\frac{n(n+2)}{8(n+1)}$
U_n^4	$\frac{1}{240}(n+2)(5n+8)$
$U_n^2 {U'_n}^2$	$\frac{1}{144}(n^2+4n+76) - \frac{2n+1}{3n(n+1)}$
$(\mathbf{T_n - \mathbb{E}(T_n)})^\mathbf{2}$	$\frac{n+2}{6} + \frac{n}{n+1}$
$(T_n - \mathbb{E}(T_n))^3$	0
$(T_n - \mathbb{E}(T_n))^4$	$\frac{1}{60}(5n^2+79n+258) - \frac{5n+2}{n(n+1)}$

	$\mathbb{E}(\cdot)$
$\mathbf{D_n}$	$\frac{n}{2}$
$\mathbf{D_n^2}$	$\frac{n+2}{12} + \frac{n^2}{4}$
$\mathbf{D_n D_n'}$	$\frac{n^2}{4} + \frac{n}{2n+2}$
D_n^3	$\frac{n(n^2+n+2)}{8}$
$D_n^2 D_n'$	$\frac{1}{24}(3n^3 + n^2 + 14n - 12) + \frac{1}{2(n+1)}$
$D_n^3 D_n'$	$\frac{1}{16}(n^4 - 4n^3 + 15n^2 - 36n + 56) - \frac{4}{(n+1)}$
D_n^4	$\frac{1}{240}(15n^4 + 30n^3 + 65n^2 + 18n + 16)$
$D_n^2 {D_n'}^2$	$\frac{1}{144}(9n^4 + 6n^3 + 85n^2 - 68n + 148) - \frac{7n+2}{6n(n+1)}$
$\mathbf{T_n^2}$	$n^2 + \frac{n+2}{6} + \frac{n}{n+1}$
T_n^3	$n^3 + \frac{n^2}{2} + 4n - 3 + \frac{3}{n+1}$
T_n^4	$n^4 + n^3 + \frac{97n^2}{12} - \frac{281n}{60} + \frac{103}{10} - \frac{11n+2}{n(n+1)}$

A.2. Type B

If $W_n = \mathtt{B}_n$, we obtain the following list of (joint) moments up to degree 4. The result for $\mathbb{E}(U_n^4)$ corresponds to Lemma 5.16 and the result for $\mathbb{E}((T_n - \mathbb{E}(T_n))^4)$ to Proposition 5.21. The moments in boldface were already known before and can be found in [KS20].

	$\mathbb{E}(\cdot)$
$\mathbf{U_n}$	0
$\mathbf{U_n^2}$	$\frac{n+1}{12}$
$\mathbf{U_n U_n'}$	$\frac{1}{4}$
U_n^3	0
$U_n^2 U_n'$	0
$U_n^3 U_n'$	$\frac{n+1}{16}$
U_n^4	$\frac{1}{240}(n+1)(5n+3)$
$U_n^2 {U_n'}^2$	$\frac{1}{144}(n^2 + 2n + 19) + \frac{2n-1}{24n(n-1)}$
$(\mathbf{T_n} - \mathbb{E}(\mathbf{T_n}))^\mathbf{2}$	$\frac{n+4}{6}$
$(T_n - \mathbb{E}(T_n))^3$	0
$(T_n - \mathbb{E}(T_n))^4$	$\frac{1}{60}(5n^2 + 39n + 79) + \frac{2n-1}{4n(n-1)}$

	$\mathbb{E}(\cdot)$
$\mathbf{D_n}$	$\frac{n}{2}$
$\mathbf{D_n^2}$	$\frac{n+1}{12} + \frac{n^2}{4}$
$\mathbf{D_n D'_n}$	$\frac{n^2+1}{4}$
D_n^3	$\frac{n(n^2+n+1)}{8}$
$D_n^2 D'_n$	$\frac{1}{24} n(7 + n + 3n^2)$
$D_n^3 D'_n$	$\frac{1}{16}(1 + n + 4n^2 + n^3 + n^4)$
D_n^4	$\frac{1}{240}(15n^4 + 30n^3 + 35n^2 + 8n + 3)$
$D_n^2 {D'_n}^2$	$\frac{1}{144}(9n^4 + 6n^3 + 43n^2 + 2n + 19) + \frac{2n-1}{24n(n-1)}$
$\mathbf{T_n^2}$	$n^2 + \frac{n+4}{6}$
T_n^3	$n(n^2 + \frac{n}{2} + 2)$
T_n^4	$n^4 + n^3 + \frac{49n^2}{12} + \frac{13n}{20} + \frac{79}{60} + \frac{2n-1}{4n(n-1)}$

Bibliography

[AFHZ14] A. C. Atkinson, V. V. Fedorov, A. M. Herzberg, and R. Zhang. Elemental information matrices and optimal experimental design for generalized regression models. *Journal of Statistical Planning and Inference*, 144:81 – 91, 2014. International Conference on Design of Experiments.

[AGS05] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005.

[AHT12] S. Aoki, H. Hara, and A. Takemura. *Gröbner Basis Techniques for Design of Experiments*, pages 261–273. Springer New York, New York, NY, 2012.

[AJLS17] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. *Information geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, 2017.

[Als09] G. Alsmeyer. *Mathematische Statistik*. Skripten zur mathematischen Statistik. Inst. für Math. Statistik, Universität Münster, third edition, 2009.

[ARS17] C. Améndola, K. Ranestad, and B. Sturmfels. Algebraic Identifiability of Gaussian Mixtures. *International Mathematics Research Notices*, 2018(21):6556–6580, 04 2017.

[BB05] A. Björner and F. Brenti. *Combinatorics of Coxeter groups*, volume 231 of *Graduate Texts in Mathematics*. Springer, New York, 2005.

[BCR13] J. Bochnak, M. Coste, and M. Roy. *Real algebraic geometry*, volume 36. Springer, New York, 2013.

[BDKS19] T. Boege, A. D'Alì, T. Kahle, and B. Sturmfels. The geometry of gaussoids. *Foundations of Computational Mathematics*, 19(4):775–812, Aug 2019.

[Bil95] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995.

[BLM+07] Y. Berstein, J. Lee, H. Maruri-Aguilar, S. Onn, E. Riccomagno, R. Weismantel, and H. P. Wynn. Nonlinear matroid optimization and experimental design. *SIAM Journal on Discrete Mathematics*, 22, 08 2007.

[BMO+10] Y. Berstein, H. Maruri-Aguilar, S. Onn, E. Riccomagno, and H. P. Wynn. Minimal average degree aberration and the state polytope for experimental designs. *Annals of the Institute of Statistical Mathematics*, 62(4):673–698, Aug 2010.

[BMS04] D. R. Berman, S. C. McLaurin, and D. D. Smith. Ranking whist players. *Discrete Mathematics*, 283(1-3):15–28, 2004.

[BR19] B. Brück and F. Röttger. A central limit theorem for the two-sided descent statistic on Coxeter groups. *arXiv e-prints*, page arXiv:1908.07955, Aug 2019.

[Bre94] F. Brenti. q-Eulerian polynomials arising from Coxeter groups. *European Journal of Combinatorics*, 15(5):417–441, 1994.

[BT52] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[CD17] S. Chatterjee and P. Diaconis. A central limit theorem for a new statistic on permutations. *Indian Journal of Pure and Applied Mathematics*, 48(4):561–573, 2017.

[Cha08] S. Chatterjee. A new method of normal approximation. *The Annals of Probability*, 36(4):1584–1610, 2008.

[Che53] H. Chernoff. Locally optimal designs for estimating parameters. *Annals of Mathematical Statistics*, 24:586–602, 1953.

[CLO15] D. A. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2015. An introduction to computational algebraic geometry and commutative algebra.

[CMP07] T. Callaghan, P. J. Mucha, and M. A. Porter. Random walker ranking for NCAA division I-A football. *American Mathematical Monthly*, 114(9):761–777, 2007.

[Cox35] H. S. M. Coxeter. The complete enumeration of finite groups of the form $R_i^2 = (R_i R_j)^{k_{ij}} = 1$. *Journal of the London Mathematical Society*, 10:21–25, 1935.

[CRS66] L. Carlitz, D. P. Roselle, and R. A. Scoville. Permutations and sequences with repetitions by number of increases. *Journal of Combinatorial Theory*, 1:350–374, 1966.

[DHM+17] C. C. Drovandi, C. Holmes, J. M. McGree, K. Mengersen, S. Richardson, and E. G. Ryan. Principles of experimental design for big data analysis. *Statistical Science*, 32(3):385–404, 08 2017.

[DMJ13] J. C. Duchi, L. Mackey, and M. I. Jordan. The asymptotics of ranking algorithms. *The Annals of Statistics*, 41(5):2292–2323, 2013.

[DMP+04] H. Dette, V. B Melas, A. Pepelyshev, et al. Optimal designs for a class of nonlinear regression models. *The Annals of Statistics*, 32(5):2142–2167, 2004.

[DS98] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26(1):363–397, 02 1998.

[DSS09] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*. Oberwolfach Seminars. Birkhäuser Basel, 2009.

[Fed72] V. V. Fedorov. *Theory of optimal experiments*. Academic Press, New York-London, 1972. Translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probability and Mathematical Statistics, No. 12.

[FH08] P. A. Freund and H. Holling. Creativity in the classroom: A multilevel analysis investigating the impact of creativity and reasoning ability on gpa. *Creativity Research Journal - CREATIVITY RES J*, 20:309–318, 08 2008.

[FHS20] F. Freise, H. Holling, and R. Schwabe. Optimal designs for two-level main effects models on a restricted design region. *Journal of Statistical Planning and Inference*, 204:45 – 54, 2020.

[Fis35] R.A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

[FK85] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[FSW17] A. F. Francis, M. Stehlík, and H. P. Wynn. "Building" exact confidence nets. *Bernoulli*, 23(4B):3145–3165, 2017.

[Fé20] V. Féray. On the central limit theorem for the two-sided descent statistics in coxeter groups. *Electronic Communications in Probability*, 25:6 pp., 2020.

[GDHS12] U. Graßhoff, A. Doebler, H. Holling, and R. Schwabe. Optimal design for linear regression models in the presence of heteroscedasticity caused by random coefficients. *Journal of Statistical Planning and Inference*, 142(5):1108–1113, 2012.

[GHRS20] U. Graßhoff, H. Holling, F. Röttger, and R. Schwabe. Optimality regions for designs in multiple linear regression models with correlated random coefficients. *Journal of Statistical Planning and Inference*, 209:267–279, 2020.

[GHS09] U. Graßhoff, H. Holling, and R. Schwabe. On optimal design for a heteroscedastic model arising from random coefficients. *Proceedings of the 6th St. Petersburg Workshop on Simulation*, 01 2009.

[GRF03] T. Graves, C. S. Reese, and M. Fitzgerald. Hierarchical models for permutations: analysis of auto racing results. *Journal of the American Statistical Association*, 98(462):282–291, 2003.

[GRRW10] P. Gibilisco, E. Riccomagno, M.P. Rogantin, and H.P. Wynn. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2010.

[GS] D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

[GS08] U. Graßhoff and R. Schwabe. Optimal design for the Bradley–Terry paired comparison model. *Statistical Methods and Applications*, 17(3):275–289, 2008.

[HT98] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

[Hun04] D. R. Hunter. MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

[Kah10] T. Kahle. *On boundaries of statistical models*. Ph.D. thesis. Leipzig University, 2010.

[Kie74] J. Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2(5):849–879, 1974.

[KKT06] K. Kobayashi, H. Kawasaki, and A. Takemura. Parallel matching for ranking all teams in a tournament. *Advances in Applied Probability*, 38(3):804–826, 2006.

[KOS16] T. Kahle, K. Oelbermann, and R. Schwabe. Algebraic geometry of Poisson regression. *Journal of Algebraic Statistics*, 7:29–44, 2016.

[KRS19] T. Kahle, F. Röttger, and R. Schwabe. The semi-algebraic geometry of optimal designs for the Bradley–Terry model. *arXiv e-prints*, page arXiv:1901.02375, Jan 2019.

[KS20] T. Kahle and C. Stump. Counting inversions and descents of random elements in finite Coxeter groups. *Mathematics of Computation*, 89(321):437–464, 2020.

[KW60] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[Leh86] E. L. Lehmann. *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, second edition, 1986.

[Leh98] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer, New York, 1998.

[Mal72] C. L. Mallows. A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43:508–515, 1972.

[MSW13] H. Maruri-Aguilar, E. Sáenz-de-Cabezón, and H. P. Wynn. Alexander duality in experimental designs. *Annals of the Institute of Statistical Mathematics*, 65(4):667–686, Aug 2013.

[Ö19] A. Y. Özdemir. Martingales and descent statistics. *arXiv e-prints*, page arXiv:1901.01719, Jan 2019.

[PAV+13] A. Papachristodoulou, J. Anderson, G. Valmorbida, S. Prajna, P. Seiler, and P. A. Parrilo. *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*. `http://arxiv.org/abs/1310.4716`, 2013. Available from `http://www.eng.ox.ac.uk/control/sostools`, `http://www.cds.caltech.edu/sostools` and `http://www.mit.edu/parrilo/sostools`.

[PB07] M. Patan and B. Bogacka. Efficient sampling windows for parameter estimation in mixed effects models. In *mODa 8—Advances in model-oriented design and analysis*, Contributions to Statistics, pages 147–155. Physica-Verlag/Springer, Heidelberg, 2007.

[Pea94] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. (A.)*, 185:71–110, 1894.

[Pet13] T. K. Petersen. Two-sided Eulerian numbers via balls in boxes. *Mathematics Magazine*, 86(3):159–176, 2013.

[Pet18] T. K. Petersen. A two-sided analogue of the Coxeter complex. *Electronic Journal of Combinatorics*, 25(4):Paper 4.64, 28, 2018.

[Pis19] G. Pistone. Information Geometry of the Probability Simplex: A Short Course. *arXiv e-prints*, page arXiv:1911.01876, Nov 2019.

[PRW00] G. Pistone, E. Riccomagno, and H.P. Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2000.

[PRW19] E. Pesce, E. Riccomagno, and H. P. Wynn. Experimental design issues in big data: The question of bias. In F. Greselin, L. Deldossi, L. Bagnato, and M. Vichi, editors, *Statistical Learning of Complex Data*, pages 193–201, Cham, 2019. Springer International Publishing.

[Puk93] F. Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1993.

[PW96] G. Pistone and H. P. Wynn. Generalised confounding with Gröbner bases. *Biometrika*, 83(3):653–666, 09 1996.

[PWZ17] L. Pronzato, H. P. Wynn, and A. A. Zhigljavsky. Extended generalised variances, with applications. *Bernoulli*, 23(4A):2617–2642, 11 2017.

[Rao45] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

[Ric09] E. Riccomagno. A short history of algebraic statistics. *Metrika*, 69(2):397–418, Mar 2009.

[RKR16] N. Rudak, S. Kuhnt, and E. Riccomagno. Numerical algebraic fan of a design for statistical model building. *Statistica Sinica*, 26, 07 2016.

[Röt20] F. Röttger. Asymptotics of a locally dependent statistic on finite reflection groups. *Electronic Journal of Combinatorics*, 27(2):P2.24, 2020.

[RS16] M. Radloff and R. Schwabe. Invariance and equivariance in experimental design for nonlinear models. In J. Kunert, C. H. Müller, and A. C. Atkinson, editors, *mODa 11 - Advances in Model-Oriented Design and Analysis: Proceedings of the 11th International Workshop in Model-Oriented Design and Analysis held in Hamminkeln, Germany, June 12-17, 2016*, pages 217–224. Springer International Publishing, Cham, 2016.

[Sha03] J. Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer, 2003.

[Sil80] S.D. Silvey. *Optimal design: an introduction to the theory for parameter estimation*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[Smi18] K. Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.

[SS89] K. R. Shah and B. K. Sinha. *Theory of optimal designs*, volume 54 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1989.

[Sul18] S. Sullivant. *Algebraic Statistics*. Graduate Studies in Mathematics. American Mathematical Society, 2018.

[SV15] C. D. Savage and M. Visontai. The $\mathbf{s}$-Eulerian polynomials have only real roots. *Transactions of the American Mathematical Society*, 367(2):1441–1466, 2015.

[SW12] B. Sturmfels and V. Welker. Commutative algebra of statistical ranking. *Journal of Algebra*, 361:264–286, 2012.

[SY99] G. Simons and Y. Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.

[Tor04] B. Torsney. Fitting Bradley Terry models using a multiplicative algorithm. In J. Antoch, editor, *COMPSTAT 2004 — Proceedings in Computational Statistics*, pages 513–526, Heidelberg, 2004. Physica-Verlag HD.

[Vat96] V. A. Vatutin. The numbers of ascending segments in a random permutation and in one inverse to it are asymptotically independent. *Diskretnaya Matematika*, 8(1):41–51, 1996.

[Vis13] M. Visontai. Some remarks on the joint distribution of descents and inverse descents. *Electronic Journal of Combinatorics*, 20(1):Paper 52, 12, 2013.

[Vog60] W. Vogel. A sequential design for the two armed bandit. *The Annals of Mathematical Statistics*, 31(2):430–443, 1960.

[Whi73] P. Whittle. Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(1):123–130, 1973.

[ZL10] S. Zhang and M. D Lee. Optimal experimental design for a class of bandit problems. *Journal of Mathematical Psychology*, 54(6):499–508, 2010.