Higher order concentration of measure and applications

Dissertation zur Erlangung des Grades

Doktor der Mathematik (Dr. math.)

vorgelegt von

Arthur Sinulis

September 2019

Universität Bielefeld

Fakultät für Mathematik

Dissertation
Higher order concentration of measure and applications

Annahme der Dissertation durch die Prüfungskommission bestehend aus:
Prof. Dr. Alexander Grigor'yan (Vorsitzender der Kommission)
Prof. Sergey Bobkov
Prof. Dr. Kai-Uwe Bux
Prof. Dr. Friedrich Götze (Betreuer)

Datum des Kolloquiums: 28. November 2019

Gedruckt auf altersbeständigem Papier gemäß ISO 9706.

Acknowledgments

Firstly I would like to thank my adviser Prof. Dr. Friedrich Götze for raising me academically from my first semester at Bielefeld University throughout to this very point. He has given me an opportunity to study such an interesting subject in a fruitful working environment and has guided me throughout my years of study. Secondly, I am very grateful to Prof. Sergey G. Bobkov for the insightful and helpful discussions. He has provided me with several ideas and references, which helped to embed this thesis in a larger framework and enhance its readability. Thirdly, I want to take this opportunity to thank Prof. Dr. Matthias Löwe for agreeing to write a report on the thesis on short notice. I imagine that the writing of a report may take a while, and I am very happy that these three distinguished mathematicians took their time to read this thesis.

Moreover, I am thankful to the four members of the board of examiners, Prof. Sergey Bobkov, Prof. Dr. Kai-Uwe Bux, Prof. Dr. Friedrich Götze and Prof. Dr. Alexander Grigor'yan (in alphabetical order), who helped me to complete my doctorate successfully.

There have been many people in this rather long time period who have supported me in various ways, and a lot of whom have become good friends throughout study, but also beyond. Namely I want to mention Hendrik Gerlach, Thoren Huppke, Christoph Kruse, Jonas Jalowy, Bastian Roser, Patrick Wegener and Sophiane Yahiatene. Furthermore, this work might not have been possible without the very fruitful mathematical relationship with Dr. Holger Sambale, who deserves a great deal of gratitude. It has always been a pleasure working with you, and I'm sure you'll be a great lecturer, adviser and researcher in the future.

The deepest thankfulness and appreciation belong to the most precious person of all the people I met, the one who has shared nearly half of her life with me, and hopefully will continue to do so for many more years to come. This work is dedicated to you, Katherina.

Abstract

In this thesis we investigate the concentration of measure phenomenon, especially with a focus on higher order concentration. The term *higher order* serves to emphasize that we are interested in multilevel concentration inequalities of a function $f(X) = f(X_1, \dots, X_n)$ of many random variables $X_1, \dots, X_n$ in terms of quantities which resemble higher order partial derivatives. More precisely, we aim for inequalities of the form

$$\mathbb{P}\left(|f(X) - \mathbb{E}\, f(X)| \geq t\right) \leq 2 \exp\Big(- c \min_{k=1,\dots,d} \frac{t^{2/k}}{C_k}\Big)$$

for all $t \geq 0$ and some constants $C_1, \dots, C_d$ depending on f, and c which might depend on the distribution of X. Up to a constant depending on d only these can be understood as extensions of Bernstein-type inequalities, i. e. we can equivalently prove concentration inequalities of the form

$$\mathbb{P}\left(|f(X) - \mathbb{E}\, f(X)| \geq t\right) \leq 2 \exp\Big(- c \frac{t^2}{C_1 + C_2 t + \sum_{k=3}^d C_k t^{2-2/k}}\Big).$$

We begin by proving Bernstein-type inequalities, i. e. the case $d = 2$, in Chapter 3. We show how the Bobkov–Götze theorem [BG99] can be used to deduce Bernstein-type inequalities for functions f with "derivatives" (in some precise sense) bounded by some sub-Gaussian random variable g. Secondly, we show how to leverage an inequality by Gao–Quastel [GQ03] to obtain concentration inequalities on the symmetric group.

To include the cases $d \geq 3$, in Chapter 4 we consider two different, but closely related frameworks. The first situation are *finite spin systems*, i. e. probability measures on finite product spaces with dependence among the coordinates. We make use of an approximate tensorization property of the entropy proven by Marton [Mar15] to establish a logarithmic Sobolev inequality, from which we deduce concentration inequalities. We demonstrate its usefulness in general Ising

models and in the exponential random graph model primarily, but also provide other examples. In the situation of independent random variables, building upon results of [BBLM05] and [BGS19], we prove analogous concentration inequalities for bounded functions.

Thereafter, in Chapter 5 we restrict ourselves to polynomials in independent random variables $X_1, \ldots, X_n$, but allow the X_i to be unbounded. This is not covered by the results in Chapter 4. However, we only deal with random variables with sub-exponentially decaying tails, i. e. $\mathbb{P}(|X_i| \geq t) \leq c \exp(-Ct^\alpha)$, and prove multilevel concentration inequalities. This generalizes results of Adamczak–Wolff [AW15] from sub-Gaussian to α-sub-exponential random variables. We complement it with several generalizations of inequalities proven in the sub-Gaussian case.

This thesis is concluded with some open questions, which could serve as starting points for future research projects.

Contents

1 Introduction 1
1.1 Concentration of measure . . . 1
1.2 Our contribution . . . 8
1.3 Discussion of related literature . . . 14
1.3.1 Concentration inequalities for general functions . . . 14
1.3.2 Concentration inequalities for polynomials . . . 19
1.3.3 Further topics . . . 22
1.4 Outline of the thesis . . . 23

2 Preliminaries 25
2.1 Notations . . . 25
2.2 Difference operators and functional inequalities . . . 27
2.3 From functional to concentration inequalities . . . 31
2.4 Independent random variables . . . 35
2.5 (Weakly dependent) Spin systems . . . 36
2.6 Examples of weakly dependent spin systems . . . 38
2.6.1 Finite product measures . . . 39
2.6.2 The Ising and the Potts model . . . 39
2.6.3 Random networks: The (vertex-weighted) exponential random graph model . . . 42
2.6.4 Models with exclusion: Random coloring and hard-core model 46

3 Bernstein-type inequalities 51
3.1 General results . . . 52
3.2 Applications . . . 56
3.2.1 Derivations . . . 56
3.2.2 Weakly dependent product measures . . . 58
3.2.3 Symmetric group . . . 60
3.2.4 Homogeneous polynomials in [0,1]-random variables . . . 65
3.3 Proofs and auxiliary results . . . 66

4 Concentration inequalities for bounded functions 73
4.1 General results . . . 73
4.2 Applications . . . 75
4.2.1 Deviation inequalities for empirical processes . . . 75
4.2.2 Concentration properties of U-statistics . . . 78
4.2.3 Polynomials in the Ising model . . . 79
4.2.4 Number of triangles in exponential random graph models . 81
4.3 Proofs . . . 84

5 Concentration inequalities for polynomials in independent random variables 93
5.1 General results . . . 95
5.2 Applications . . . 97
5.2.1 Euclidean norm of a vector with independent components . 97
5.2.2 Projections and distance to a fixed subspace . . . 99
5.2.3 Spectral bound for a product of a fixed and a random matrix 100
5.2.4 Concentration properties for fixed design linear regression . 101
5.2.5 Special cases . . . 102
5.3 The multilinear case: Proof of Theorem 5.4 . . . 102
5.4 Hanson–Wright-type inequality: Proof of Proposition 5.5 105
5.5 The polynomial case: Proof of Theorem 5.6 . . . 109
5.6 The general sub-exponential case: $\alpha \in (0,1]$ . . . 115

A Approximate tensorization of entropy in finite spaces 123

B Properties of Orlicz quasinorms 127

C LSIs and difference operators 131

Open questions 135

Bibliography 139

CHAPTER 1

Introduction

Since the inception of probability theory as a mathematical subject, a driving force of its development has been the desire to understand the behavior of functions of independent random variables. To illustrate the transition from the arguably most famous theorems in probability theory to the content of this thesis, let us consider the simple random walk given by $S_n = \sum_{i=1}^n X_i$ for independent, identically distributed (i. i. d.) centered random variables $X_1, \ldots, X_n$. The *law of large numbers* (LLN) ensures the convergence $n^{-1} S_n \to 0$ either in probability or almost surely, depending on the integrability conditions of the summands. However, as the theorem is very general, it is unsatisfactory due to two disadvantages. Firstly, it is an asymptotic result, i. e. it does not provide any information for a fixed $n \in \mathbb{N}$ and thus no rigorous results for finite sample sizes. And secondly, it does neither give any information on the fluctuations, nor any hint at the scale of fluctuation of S_n around its mean.

A sharpening of the LLN is given by the *central limit theorem* (CLT), arguably the single most important result in probability theory. Applied to the random walk S_n, it states that under under a finite second moment of the X_i the random variable $n^{-1/2} S_n$ tends to a non-deterministic limit, the normal distribution. Equivalently, the scale at which one can see stochastic fluctuations is $\sqrt{n}$ instead of n, i. e. at the scale of the standard deviation. This does solve the second issue, but not the first one, as the CLT is also an asymptotic result. These issues lead to the investigation of properties of finite sample sizes.

1.1 Concentration of measure

There are now two leading questions which arise:

(a) Is it possible to bound the tails of S_n for finite n? If we expect the fluctuations to be of order $\sqrt{n}$, can we provide upper bounds on its tail behavior?

(b) What can be said about more general functions $f(X_1, \ldots, X_n)$ and their concentration properties? Is the special structure of the sum important or is it possible to provide similar descriptions for a wider class of functions, i. e. the probability $\mathbb{P}(|S_n - \mathbb{E}\, S_n| \geq t)$ for some $t > 0$?

Obviously the first question is too general in the class of all random variables, even under a finite variance assumption. It is reasonable to consider random variables that themselves have light tails, i. e. for which the tails decay sufficiently rapidly.

There are many possible routes one can take to shed some light on these questions. In the case of S_n, a straightforward approach is applying Markov's inequality to the function $x \mapsto \exp(\lambda x)$ for any $\lambda > 0$ and optimization. For example, assuming that all the random variables X_i are sub-Gaussian with constant σ^2 in the sense that for all $\lambda \in \mathbb{R}$ it holds

$$\mathbb{E} \exp(\lambda X_i) \leq \exp\Big(\frac{\sigma^2 \lambda^2}{2}\Big), \tag{1.1}$$

we obtain by straightforward calculations

$$\mathbb{P}\big(S_n \geq t\big) \leq \exp\Big(-\frac{t^2}{2\sigma^2 n}\Big). \tag{1.2}$$

As $-X_i$ also satisfies (1.1), by a union bound this easily yields

$$\mathbb{P}\big(|S_n| \geq t\big) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2 n}\Big). \tag{1.3}$$

The implications of (1.3) are far-reaching. Indeed, for any (failure probability) $\varepsilon \in (0,1)$ the inequality implies that with probability at least $1-\varepsilon$ we have

$$|S_n| \leq \sqrt{2\sigma^2 n \log(2/\varepsilon)}.$$

Here, the $\sqrt{n}$ scale also appears naturally, as (up to a set of probability less than ε) S_n only takes values of order $\sqrt{n}$. Furthermore, if we want to increase the probability that the inequality holds, i. e. decrease ε, we only pay a logarithmic price for that. A naive application of Chebyshev's inequality would result in a bound of the form $\sqrt{\sigma^2 n/\varepsilon}$.

The *deviation inequality* (1.2) and the *concentration inequality* (1.3) are instances of what can be called concentration of measure, as it gives non-asymptotic bounds on deviations of a random variable from a deterministic value (in this case, the mean). However, the big drawback of this approach is its strong reliance on the function S_n being linear, allowing to factorize the moment generating function and use the sub-Gaussian property of its components.

A more successful way to prove concentration of measure is by using functional inequalities, or more specifically an approach which is nowadays known as the *entropy method*. It has emerged as a way to prove several groundbreaking concentration inequalities in product spaces by Talagrand [Tal91a; Tal96b], mainly in the papers of Ledoux [Led97], Bobkov and Ledoux [BL97], and Massart [Mas00]. Before explaining it in the general setting, let us recall the *logarithmic Sobolev inequality* (LSI) for the standard Gaussian measure γ on $\mathbb{R}^n$. In [Gro75], Gross established an inequality for γ which is able to capture (and actually is equivalent to) hypercontractivity results of Nelson [Nel73], which he coined the logarithmic

Sobolev inequality. It states that for any $f \in C_c^\infty(\mathbb{R}^n)$ the inequality

$$\mathrm{Ent}_\gamma(f^2) \le 2 \int |\nabla f|^2 d\gamma \tag{1.4}$$

holds, where

$$\mathrm{Ent}_\gamma(f^2) = \int f^2 \log f^2 d\gamma - \int f^2 d\gamma \log \int f^2 d\gamma$$

is the *entropy functional* (also known as Boltzmann–Shannon entropy, or Kullback–Leibler divergence). Informally, (1.4) bounds the disorder of a function f (under γ) by its average local fluctuations, measured in terms of the length of the gradient. Actually, the LSI is a very strong inequality, as it implies the famous Gaussian Poincaré inequality as proven by Chernoff [Che81] (see also [Kla85]), which itself implies that Lipschitz functions have variance at most 1 with respect to the Gaussian measure.

Furthermore, the inequality (1.4) can be used to prove concentration of measure for Lipschitz functions. One possible way is known as *Herbst argument* and we briefly sketch it. In essence, it provides a differential inequality for the moment generating function. Assume that f is bounded, smooth and L-Lipschitz for some $L > 0$ and let $g := \exp(\lambda f/2)$. Applying (1.4) to g yields

$$\mathrm{Ent}_\gamma(g^2) \le 2 \int |\nabla g|^2 d\gamma = \frac{\lambda^2}{2} \int |\nabla f|^2 e^{\lambda f} d\gamma \le \frac{\lambda^2 L^2}{2} \int e^{\lambda f} d\gamma.$$

If we denote by $H(\lambda) := \int e^{\lambda f} d\gamma$ the moment generating function (mgf) of f, this inequality can be written as

$$(\lambda^{-1} \log H(\lambda))' \le \frac{L^2}{2}.$$

By elementary tools, it can be seen that $\lambda^{-1} \log H(\lambda) \to \int f d\gamma$ as $\lambda \to 0$, so that the differential inequality can be integrated to prove

$$\log H(\lambda) \le \frac{\lambda^2 L^2}{2} + \lambda \int f d\gamma.$$

Thus, the growth of the mgf can be bounded, and Markov's inequality provides deviation inequalities of the form $\gamma(f - \int f d\gamma \ge t) \le \exp(-t^2/(2L^2))$. Repeating the argument for $-f$ yields a sub-Gaussian bound

$$\gamma\Big(|f - \int f d\gamma| \ge t\Big) \le 2 \exp\Big(-\frac{t^2}{2L^2}\Big). \tag{1.5}$$

A remarkable feature of (1.5) is the fact that the right hand side does not depend on n, which is termed *dimension-free concentration inequality*. Moreover, it is easily seen that (1.5) is a more general form of (1.3), as $f(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$

satisfies $|\nabla f| = \sqrt{n}$. Finally, note that this approach does not use any specific properties of the Gaussian measure γ, so that this argument is valid in a broader range. Indeed, whenever *any* probability measure ν satisfies the inequality (1.4), Lipschitz functions have sub-Gaussian tail decay under ν. We say that ν satisfies an LSI(σ^2) if the constant 2 in (1.4) is replaced by $2\sigma^2$. Hence we have sketched the proof of the following well-known theorem.

Theorem 1.1. *Assume that a probability measure ν satisfies an* LSI(σ^2). *For any smooth Lipschitz function f with $|\nabla f| \leq L$ and any $t \geq 0$ it holds*

$$\nu\Big(|f - \int f d\nu| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2 L^2}\Big).$$

Comparing this to the concentration properties of the random walk S_n, we also have that for any $\varepsilon \in (0,1)$, with probability at least $1-\varepsilon$ we have

$$|f - \int f d\nu| \leq \sqrt{2\sigma^2 L^2 \log(2/\varepsilon)}.$$

After having observed some prime examples of the concentration of measure phenomenon, let us provide a famous description of what it actually is. To paraphrase Talagrand [Tal96a], it can be described as the phenomenon that any function $f = f(X_1, \dots, X_n)$ of independent random variables $X_1, \dots, X_n$ tends to be very close to a deterministic quantity[1], if it is not too sensitive to one of its parameters. The function is usually assumed to be Lipschitz in some sense, depending on a suitably adapted notion of a gradient. In other words, the distribution of any Lipschitz function of independent random variables shows strong (more precisely: sub-Gaussian) concentration properties. Since it is most successful in high dimensions, it is best described as a high-dimensional phenomenon.

Theorem 1.1 is quite satisfactory, as it provides bounds on the fluctuations and is yet general enough to cover the class of Lipschitz functions. Nevertheless, it falls short of providing estimates for non-Lipschitz functions, such as polynomials. Of course there is a perfectly good reason for that, as non-Lipschitz functions need not satisfy sub-Gaussian estimates. For example, if X has a standard normal distribution, then X^2 has a $\chi^2(1)$ distribution, which has exponentially decaying tails. However, before we begin presenting concentration of measure results for non-Lipschitz functions in Sections 1.2 and 1.3, let us remark on the "classical" case described so far.

Firstly, to make sense of the term logarithmic Sobolev inequality, it is helpful draw a connection to the theory of partial differential equations. In PDE theory, a Sobolev inequality is any type of inequality that compares some norm of a function f with some (possibly different) norm of its gradient ∇f[2]. For example,

[1] This is usually its expected value or its median.

[2] Or, more generally, its weak derivative. For our purposes, we always think of smooth functions.

both norms could be L^p or L^q norms with respect to the Lebesgue measure. Let us write $\|f\|_p$ for the L^p norm of a measurable function with respect to the Lebesgue measure on $\mathbb{R}^n$. The celebrated *Gagliardo–Nirenberg–Sobolev inequality* states that for any $p \in [1,n)$ (with its conjugate $(p^*)^{-1} = p^{-1} - n^{-1}$) there is a constant $C = C(n,p)$ such that the inequality

$$\|f\|_{L^{p^*}} \le C\||\nabla f|\|_{L^p}$$

holds for any $f \in C_c^\infty(\mathbb{R}^n)$. In the case $p = 2$ and $p^* = 2n/(n-2) > 2$ we can deduce $f \in L^{p^*}$ from $|\nabla f| \in L^2$. However, there are two disadvantages: the constant C depends the dimension n, and for $n \to \infty$ there is hardly any information, i. e. it basically reduces to the *Poincaré inequality*

$$\int \Big(f - \int f dx\Big)^2 dx \le C \int |\nabla f|^2 dx.$$

In contrast, the constant 2 in (1.4) is independent of the dimension, but it only provides a weaker estimate. Indeed, if $|\nabla f| \in L^2(\gamma)$, then f is contained in the Orlicz space

$$L^2 \log L(\gamma) = \Big\{g : \mathbb{R}^n \to \mathbb{R} : \int g^2 \log(g^2) d\gamma < \infty\Big\}.$$

Note that for any probability measure ν the (strict) inclusion $L^{2+\varepsilon}(\nu) \subset L^2 \log L(\nu)$ holds for any $\varepsilon > 0$. Consequently, the logarithmic Sobolev inequality merely implies an improvement of the integrability of f by a logarithmic order, when compared to the Poincaré inequality.

Secondly, aside from the approach outlined above there is a second ansatz to deduce concentration inequalities from (1.4). We want to stress that the Herbst argument is easier to digest for a first reading, and thus we have chosen to present it, although the second approach is closer to the content of this thesis. It is possible to derive from (1.4) L^p norm inequalities for $p \ge 2$ of the type

$$\|f - \mathbb{E} f\|_p \le \sqrt{2(p-1)}\||\nabla f|\|_p. \tag{1.6}$$

Such inequalities were initially proven by Aida and Stroock [AS94] and can similarly be used to prove sub-Gaussian concentration under the condition $|\nabla f| \le 1$.[1] Furthermore, the Aida–Stroock approach has the added benefit that (1.6) allows for estimates going beyond the setting of Lipschitz functions, as it compares the L^p norm of the function under consideration to the L^p norm of its gradient. Now, it is possible to iterate this by considering the function $g := |\nabla f|$. As this thesis will make use of inequalities of the type (1.6) and an iteration procedure, we omit the details for now.

Thirdly, concentration of measure has become an established part of probability

[1] One needs to recall the classical fact that the sub-Gaussian property is equivalent to $\sup_{p\ge1} p^{-1/2}\|X\|_p < \infty$, see e. g. [Ver18, Proposition 2.5.2].

theory within the last 40 years with applications in numerous fields, as is witnessed by the monographs [MS86], [Led01], [BLM13], [RS14], [Han16]. The focus of this exposition shall not be the concentration of measure phenomenon for Lipschitz-type functions, as this is by now a rather classical topic and many excellent descriptions can be found in the above mentioned sources or in the PhD thesis of Chatterjee [Cha05, Chapter 2]. Nevertheless, we believe it is appropriate to briefly mention some landmark works.

The first result that could be considered within the framework of concentration of measure is Lévy's isoperimetric inequality on the sphere, stating that the solutions of the isoperimetric problem on the sphere are spherical caps[1] (see e. g. [MS86, Appendix I][2]). It is fair to assume that Lévy did not aim for concentration of measure results, but for a solution of the (very classical) isoperimetric problem. Hence, the birth of the theory is sometimes attributed to V. Milman[3], who reproved a famous result on almost Euclidean subspaces of high-dimensional spaces in [Mil71], for which we again refer to the monograph [MS86]. Some years later, Borell [Bor75] and Ibragimov, Sudakov and Tsirelson [CIS76; ST74] proved similar inequalities in the theory of Gaussian processes, which helped popularizing the concept. Afterwards it was further developed in a series of remarkable works in the nineties by Talagrand, proving a variety of different isoperimetric and concentration inequalities on general product spaces in [Tal88; Tal91a; Tal91b; Tal95; Tal96a; Tal96b]. As mentioned above, some of the results were reproven using the entropy method in [BL97; Led97], and the method was developed in [Mas00] and [BLM03] for general functions of independent random variables.

The reader might be inclined to ask whether the concentration of measure phenomenon is fully characterized by an inequality similar to (1.5), or whether there are generalizations of it. We argue that the results considered so far are *first order* concentration of measure in the sense that we have control of a first order difference (the gradient), leading to sub-Gaussian estimates. In contrast, the term *higher order* concentration of measure shall emphasize that we are concerned with functions which are not of Lipschitz type, and in turn obtain *multilevel concentration inequalities* which are of the form

$$\mathbb{P}(|f(X) - \mathbb{E}\, f(X)| \geq t) \leq 2 \exp\Big(- \min_{k=1,\ldots,d} \frac{t^{2/k}}{C_k} \Big)$$

for some *order* d.

The problem of applying results leading to sub-Gaussian estimates in the wrong situation can be seen in the following toy example. A handy and easy-to-use tool to prove concentration of measure is the so-called *bounded differences inequality*. It states that for any function $f : \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \to \mathbb{R}$ satisfying $|f(x) - f(y)| \leq c_i$ whenever x and y differ in the i-th coordinate only, and independent random

[1] Spherical caps are balls with respect to the geodesic distance.

[2] Milman also proved a generalization of an inequality of Lévy for manifolds with positive Ricci curvature.

[3] For example by Michel Talagrand, see [Tal91a; Tal96b].

variables X_i with values in $\mathcal{X}_i$, we have

$$\mathbb{P}\left(|f(X_1,\ldots,X_n) - \mathbb{E}\, f(X_1,\ldots,X_n)| \geq t\right) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big). \tag{1.7}$$

It was originally formulated and proven by McDiarmid [McD89], although it is a simple application of Azuma's inequality [Azu67]. Now, if we consider the function $f(x) = x_1 \sum_{j=2}^n x_j$ and $[-1,+1]$-valued, centered i. i. d. random variables $X_1,\ldots,X_n$, a short calculation shows $c_1 = 2(n-1)$ and $c_2 = \ldots = c_n = 2$. Thus, (1.7) yields

$$\mathbb{P}\left(|f(X_1,\ldots,X_n) - \mathbb{E}\, f(X_1,\ldots,X_n)| \geq t\right) \leq 2\exp\Big(-\frac{t^2}{2(n-1)^2 + 2(n-1)}\Big).$$

The right hand side does not vanish as $n \to \infty$ if we choose t of order n, which is equivalent to multiplying f by n^{-1}. However, the variance of $f(X_1,\ldots,X_n)$ is given by $\mathrm{Var}(X_1)^2(n-1)$, and so a normalization of n^{-1} is of the wrong order; in this case, by Chebyshev's inequality we obtain for any $t > 0$

$$\mathbb{P}\left(n^{-1}|f(X_1,\ldots,X_n) - \mathbb{E}\, f(X_1,\ldots,X_n)| \geq t\right) \leq \frac{\mathrm{Var}(X_1)}{nt^2} \to 0.$$

The wrong order in the bounded differences inequality is due to the fact that we need to uniformly bound the function $g_1(X) = |\sum_{j=2}^n X_j|$, and the only possible bound is of order n. On the other hand, the variance of $g_1(X)$ is of order n as well, so that a normalization of $n^{-1/2}$ should suffice. In this example, one can find the correct normalization by considering the following well-known *Hanson–Wright inequality*, which was originally proven in [HW71; Wri73], with a modern proof given in [RV13].

Theorem 1.2. *Consider independent, centered random variables $X_1,\ldots,X_n$ which are sub-Gaussian in the sense of* (1.1) *for some constant K, let $A = (a_{ij})$ be a symmetric matrix and define $f(X) = f(X_1,\ldots,X_n) = \sum_{i,j=1}^n a_{ij}X_iX_j$. There is a universal constant $c > 0$ such that for any $t \geq 0$ it holds*

$$\mathbb{P}\left(|f(X) - \mathbb{E}\, f(X)| \geq t\right) \leq 2\exp\Big(-c\min\Big(\frac{t^2}{K^4\|A\|_{\mathrm{HS}}^2}, \frac{t}{K^2\|A\|_{\mathrm{op}}}\Big)\Big). \tag{1.8}$$

Here, $\|A\|_{\mathrm{HS}} = (\sum_{ij} a_{ij}^2)^{1/2}$ is the Hilbert–Schmidt norm and $\|A\|_{\mathrm{op}} = \sup_{|x|\leq 1}|Ax|$ the operator norm of A (with respect to the Euclidean norm $|\cdot|$).

For the above example, we can choose the matrix A to be 1 for all indices (i,j) such that either $i = 1$ or $j = 1$, and calculate $\|A\|_{\mathrm{HS}} \sim \sqrt{n}$ and $\|A\|_{\mathrm{op}} \sim \sqrt{n}$.[1] Nevertheless, the Hanson–Wright inequality is only valid for quadratic forms, and cannot be applied to d-homogeneous forms such as $f(x) = x_1x_2\sum_{j=3}^n x_j$, which necessitates different concentration inequalities.

[1]The sub-Gaussian property with $K = 1$ follows from Hoeffding's lemma.

We want to highlight two main differences between the sub-Gaussian bound (1.5) and the bound in the Hanson–Wright inequality (1.8). Firstly, the scales at which the quadratic form fluctuates are determined by two norms of the matrix A, instead of the supremum of the gradient. And secondly, there are multiple levels of decay; for large values of t (more precisely, for $t \geq K^2 \|A\|_{\mathrm{HS}}^2 \|A\|_{\mathrm{op}}^{-1}$) we have an exponential tail decay instead of a Gaussian one.

1.2 Our contribution

First, we want to mention the papers from which this thesis draws most of its results.

- [GSS19b]: Friedrich Götze, Holger Sambale and Arthur Sinulis. 'Higher order concentration for functions of weakly dependent random variables'. (published in *Electron. J. Probab.*)
- [SS18]: Holger Sambale and Arthur Sinulis. 'Logarithmic Sobolev inequalities for finite spin systems and applications'. (accepted for publication in *Bernoulli*)
- [GSS18]: Friedrich Götze, Holger Sambale and Arthur Sinulis. 'Concentration inequalities for bounded functionals via generalized log-Sobolev inequalities'. (submitted)
- [GSS19a]: Friedrich Götze, Holger Sambale and Arthur Sinulis. 'Concentration inequalities for polynomials in α-sub-exponential random variables'. (submitted)
- [SS19]: Holger Sambale and Arthur Sinulis. 'Modified log-Sobolev inequalities and two-level concentration'. (submitted)

Chapter 2 contains mostly technical results, and these haven been proven in [GSS19b] and [SS18]. The third Chapter is based on [SS19], although some applications to the symmetric group as well as the concentration inequalities for polynomials in [0,1] random variables do not appear therein. We have implemented results from [SS18] and [GSS18] into Chapter 4. Our last chapter, Chapter 5, is completely based on the article [GSS19a].

Now let us provide a panoramic view of some of the results we prove in this thesis. The first concentration inequality that will be proven in Chapter 3 is the following theorem. We refer to Chapter 3 for the definitions of a difference operator and the $\Gamma-\mathrm{mLSI}(\rho)$, which are both taken from [BG99]. At first reading, one can think of the Euclidean length of the gradient, $\Gamma(f) = |\nabla f|$, as a difference operator whenever the probability measure μ is defined on $\mathbb{R}^n$.

Theorem 3.1. *Assume that a probability measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$ satisfies a $\Gamma-\mathrm{mLSI}(\rho)$ for some difference operator Γ and $\rho > 0$. Let $f, g : \mathcal{X} \to \mathbb{R}$*

be two measurable functions such that $\Gamma(f) \le g$, and g fulfills $\Gamma(g) \le b$ for some constant $b > 0$. Then for all $t \ge 0$ the inequality

$$\mu\big(f - \mathbb{E}_\mu f \ge t\big) \le \frac{4}{3} \exp\Big(- \frac{1}{8\rho} \min\Big(\frac{t^2}{(\mathbb{E}_\mu g)^2}, \frac{t}{b} \Big)\Big).$$

holds. If moreover $\Gamma(\lambda f) = |\lambda| \Gamma(f)$ for all $\lambda \in \mathbb{R}$, then for all $t \ge 0$ we have

$$\mu\big(|f - \mathbb{E}_\mu f| \ge t\big) \le 2 \exp\Big(- \frac{1}{12\rho} \min\Big(\frac{t^2}{(\mathbb{E}_\mu g)^2}, \frac{t}{b} \Big)\Big).$$

Informally, Theorem 3.1 states that under the modified logarithmic Sobolev condition one can always control the right tail of a function f by its majorizing function g. If the difference operator also satisfies an additional property, then the left tail can also be controlled, resulting in concentration inequalities.

To get a better understanding of Theorem 3.1, we present two examples. The first example is the n-sphere $S^{n-1} := \{x \in \mathbb{R}^n : |x| = 1\}$, where $|x|$ is the Euclidean norm. It has been known since the contribution of Mueller and Weissler [MW82] that S^{n-1} satisfies a logarithmic Sobolev inequality with respect to the spherical gradient (see also [BCG17]), which leads to the following application.

Proposition 3.7. *Consider S^{n-1} equipped with the uniform measure σ_{n-1} and let $f : S^{n-1} \to \mathbb{R}$ be a C^2 function satisfying $\sup_{\theta \in S^{n-1}} \|f''_S(\theta)\|_{\mathrm{op}} \le 1$. For any $t \ge 0$ it holds*

$$\sigma_{n-1}\big(|f - \mathbb{E}_{\sigma_{n-1}} f| \ge t\big) \le 2 \exp\Big(- \frac{(n-1)}{12} \min\Big(\frac{t^2}{(\mathbb{E}_{\sigma_{n-1}} |\nabla_S f|)^2}, t \Big)\Big).$$

In particular, if f is orthogonal to all affine functions, then

$$\sigma_{n-1}\big((n-1)|f - \mathbb{E}_{\sigma_{n-1}} f| \ge t\big) \le 2 \exp\Big(- \frac{1}{12} \min\Big(\frac{t^2}{\mathbb{E}_{\sigma_{n-1}} \|f''_S\|^2_{\mathrm{HS}}}, t \Big)\Big).$$

This proposition shows that functions $f : S^{n-1} \to \mathbb{R}$ which are orthogonal to all affine functions are strongly concentrated. In particular, given any $\varepsilon \in (0,1)$, with probability at least $1 - \varepsilon$ (with respect to the uniform distribution) it holds

$$|f - \mathbb{E}_{\sigma_{n-1}} f| \le (n-1)^{-1} \max\Big(\sqrt{12\, \mathbb{E} \|f''_S\|^2_{\mathrm{HS}} \log(2/\varepsilon)}, 12 \log(2/\varepsilon) \Big).$$

Our second example is the symmetric group S_n. We let τ_{ij} denote the transposition of the i-th and the j-th element. A consequence of a result in [GQ03] is the following proposition.

Proposition 3.14. *Let $n \in \mathbb{N}$, S_n be the symmetric group and π_n be the uniform distribution on S_n. Define the difference operator Γ via*

$$\Gamma(f)(\sigma) = \Big(n^{-1} \sum_{i,j} (f(\sigma) - f(\sigma \tau_{ij}))^2 \Big)^{1/2} \quad \textit{for} \quad f : S_n \to \mathbb{R}\,.$$

Then a $\Gamma-\mathrm{mLSI}(1)$ *holds.*

Proposition 3.14 makes it possible to apply Theorem 3.1 to the symmetric group and obtain concentration inequalities. For example, in Chapter 3 we provide results for Lipschitz functions with respect to different metrics. An easy corollary of Proposition 3.14 is the following concentration inequality.

Corollary. *Let* S_n *be the symmetric group,* π_n *be the uniform measure and* $F(\sigma) = \sum_{i=1}^n \mathbb{1}_{\sigma(i)=i}$ *be the number of fixed points. For any* $t \geq 0$ *it holds*

$$\pi_n(|F-1| \geq t) \leq 2\exp\Big(-\frac{t^2}{32+8t/3}\Big).$$

In particular, this implies that for any $\varepsilon \in (0,1)$, with probability at least $(1-\varepsilon)$ (which amounts to saying that for at least $(1-\varepsilon)\cdot n!$ permutations) we have

$$|F-1| \leq \frac{16}{3}\log(2/\varepsilon) + \sqrt{32\log(2/\varepsilon)}.$$

The asymptotic distribution of the number of fixed points is known to be a Poisson distribution, so that for large values of t we lose a logarithmic factor in t in the exponential. On the other hand, this result is non-asymptotic, i. e. it can be applied to any $n \in \mathbb{N}$. Similar results have been derived using different methods in [Cha05].

Furthermore, Chapter 3 also contains a proof of a slightly weaker version of the famous convex distance inequality by Talagrand for the symmetric group. As far as we aware, there has been no proof of either the weak of the strong version in this setting using the entropy method. We defer the definition of the convex distance to Chapter 3.

Proposition 3.3. *Let* S_n *be the symmetric group and* π_n *be the uniform distribution on* S_n. *For any set* $A \subseteq S_n$ *with* $\pi_n(A) \geq 1/2$ *and all* $t \geq 0$ *we have*

$$\pi_n(d_T(\cdot,A) \geq t) \leq 2\exp\Big(-\frac{t^2}{64}\Big).$$

Chapter 4 contains concentration inequalities for probability measures satisfying a specific logarithmic Sobolev inequality. Here, we refrain from giving the necessary definitions and the general concentration result (Theorem 4.1 and Proposition 4.3), and rather concentrate on some applications. The first one is a uniform version of the Hanson–Wright inequality for the *Curie–Weiss model*, which we briefly describe. Given a parameter $\beta \geq 0$ (physically, up to a multiplicative constant, the inverse temperature), it is the probability measure on $\{-1,+1\}^n$ given by

$$\mu_\beta(x) = Z^{-1}\exp\Big(\frac{\beta}{2n}\sum_{i\neq j} x_i x_j\Big),$$

where Z is the normalization constant (also known as *partition function*). Applied to this model, we obtain the following theorem.

Theorem 4.4. *Let μ_β be the Curie–Weiss model with parameter $\beta \in (0,1)$, and fix some set $\mathcal{T}$ with elements of the form $t = (t_{i_1 \dots i_d})_{1 \le i_1 < \dots < i_d \le n}, t_{i_1 \dots i_d} \in \mathbb{R}$. Set*

$$f(x) \coloneqq f_{\mathcal{T}}(x) \coloneqq \sup_{t \in \mathcal{T}} \Big| \sum_{i_1 < \dots < i_d} t_{i_1 \dots i_d} x_{i_1} \cdots x_{i_d} \Big|.$$

There are d functions $W_1, \dots, W_d$ (for the exact definition, see (4.7)) and a constant $\sigma^2 = \sigma^2(\beta)$, which is independent of $n \in \mathbb{N}$, such that for any $t \ge 0$ it holds

$$\mu_\beta \left(f(x) - \mathbb{E}_{\mu_\beta} f \ge t \right) \le e \exp\Big(- \frac{1}{4\sigma^2} \min_{k=1,\dots,d} \Big(\frac{t}{de \, \mathbb{E}\, W_k} \Big)^{2/k} \Big).$$

In essence, Theorem 4.4 provides the counterpart of some known concentration inequalities for independent Rademacher random variables (such as the ones in [BBLM05]) to the case of dependent random variables. The non-uniform case can also be treated, and we even prove concentration inequalities for such multilinear functions.

Theorem 4.9. *Let μ_β be the Curie–Weiss model with temperature $\beta \in (0,1)$. Let $A = (a_{i_1,\dots,i_d})$ be a symmetric d-tensor with vanishing diagonal and bounded entries $\sup_{i_1,\dots,i_d} |a_{i_1 \dots i_d}| = 1$, and define $f : \{-1,+1\}^n \to \mathbb{R}$ via*

$$f \coloneqq f(x) \coloneqq \sum_{I=(i_1,\dots,i_d)} a_I x_{i_1} \cdots x_{i_d}.$$

There is a constant $\sigma^2 = \sigma^2(\beta)$ such that for all $t \ge 0$ it holds

$$\mu_\beta\big(|f(x) - \mathbb{E}_{\mu_\beta} f| \ge t \big) \le 2 \exp\Big(- \frac{1}{\sigma^2 n} \min(t^2, t^{2/d}) \Big).$$

In particular, this theorem shows that the fluctuations of d-forms with bounded coefficients are of order $n^{d/2}$, as by rescaling we obtain

$$\mu_\beta(|f(x) - \mathbb{E}_{\mu_\beta} f| \ge n^{d/2} t) \le 2 \exp\Big(- \frac{1}{\sigma^2} \min(n^{d-2} t^2, t^{2/d}) \Big).$$

As the maximal value of f could be of order n, this shows that such functions are strongly concentrated with respect to μ_β. Consequently, for $\beta \in (0,1)$ the dependence between the spins that is introduced in the Curie–Weiss model does not lead to significantly different behavior than in the independent case. This is in good accordance with limit theorems in this regime, see e. g. [EN78], [Ell06], [EL10]. Theorem 4.9 has been generalized by Adamczak and co-workers [AKPS19].

Another application is counting triangles in an *exponential random graph model*. For the exact definition of this model, see Section 2.6. It is a model of a random graph which incorporates dependencies among the (random) connections between the vertices. We let $\binom{\mathcal{I}_n}{3}$ be the set of all possibilities of choosing three distinct edges and

$$\mathcal{T}_n := \Big\{\{e,f,g\} \in \binom{\mathcal{I}_n}{3} : e,f,g \text{ form a triangle}\Big\}.$$

The statistic we are interested in is the number of triangles

$$T_3(x) := \sum_{\{e_1,e_2,e_3\} \in \mathcal{T}_n} x_{e_1} x_{e_2} x_{e_3}.$$

Furthermore, we let $f_1(x) := \sum_{e \in \mathcal{I}_n} (x_e - \mathbb{E}_{\mu_\beta}(x_e))$ be the number of edges in such a graph, and $\mu_2 := \mathbb{E}_{\mu_\beta} x_e x_f$ (for some edges $e \neq f \in \mathcal{I}_n$, $e \cap f \neq \emptyset$).

Theorem 4.11. *Under certain technical conditions on the exponential random graph model μ_β, there exists a constant $C > 0$ such that for all $t \geq 0$ we have the multilevel concentration bounds*

$$\mu_\beta(|T_3 - \mathbb{E}_{\mu_\beta} T_3| \geq t) \leq 2\exp\Big(-\frac{1}{C}\min\Big(\Big(\frac{t}{n^{3/2}}\Big)^{2/3}, \frac{t}{\mu_1 n^{3/2}}, \Big(\frac{t}{\mu_2 n^2}\Big)^2\Big)\Big)$$
$$\mu_\beta(|T_3 - \mathbb{E}_{\mu_\beta} T_3 - (n-2)\mu_2 f_1| \geq t) \leq 2\exp\Big(-\frac{1}{C}\min\Big(\Big(\frac{t}{n^{3/2}}\Big)^{2/3}, \frac{t}{\mu_1 n^{3/2}}\Big)\Big).$$

In Chapter 4 we give a detailed discussion of this result, but for this expository outline we restrict to the following histograms.

As a corollary, it is possible to deduce limit theorems for the number of triangles from those of the number of edges in such models. Clearly one has to still check whether there is a CLT for the edge count, but this is a much more tractable random variable, so that the corollary reduces the problem to a "linear" one.

Corollary 4.12. *Let μ_β be as above. Assuming $\binom{n}{2}^{-1/2} \sum_{e \in \mathcal{I}_n} (x_e - \mathbb{E}_{\mu_\beta} x_e) \Rightarrow \mathcal{N}(0,v^2)$, we can infer*

$$\frac{T_3 - \mathbb{E}_{\mu_\beta} T_3}{(n-2)\mu_2 \sqrt{\binom{n}{2}}} \Rightarrow \mathcal{N}(0,v^2).$$

Lastly, in Chapter 5 we prove concentration inequalities for polynomials in independent random variables $X_1, \ldots, X_n$ which have α-sub-exponential tails. By this, we mean that there are two constants $c, C > 0$ such that for all $t \geq 0$ it holds

$$\mathbb{P}(|X_i| \geq t) \leq c\exp\Big(-\frac{t^\alpha}{C}\Big).$$

A simplified version of one of the results in Chapter 5 gives two-level concentration inequalities for such random variables, which are akin to the Hanson–Wright inequality (1.8).

Theorem 5.1. *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbb{E} X_i = 0, \mathbb{E} X_i^2 = 1$ and which are α-sub-exponential with constant M for $\alpha \in (0,1] \cup \{2\}$.*

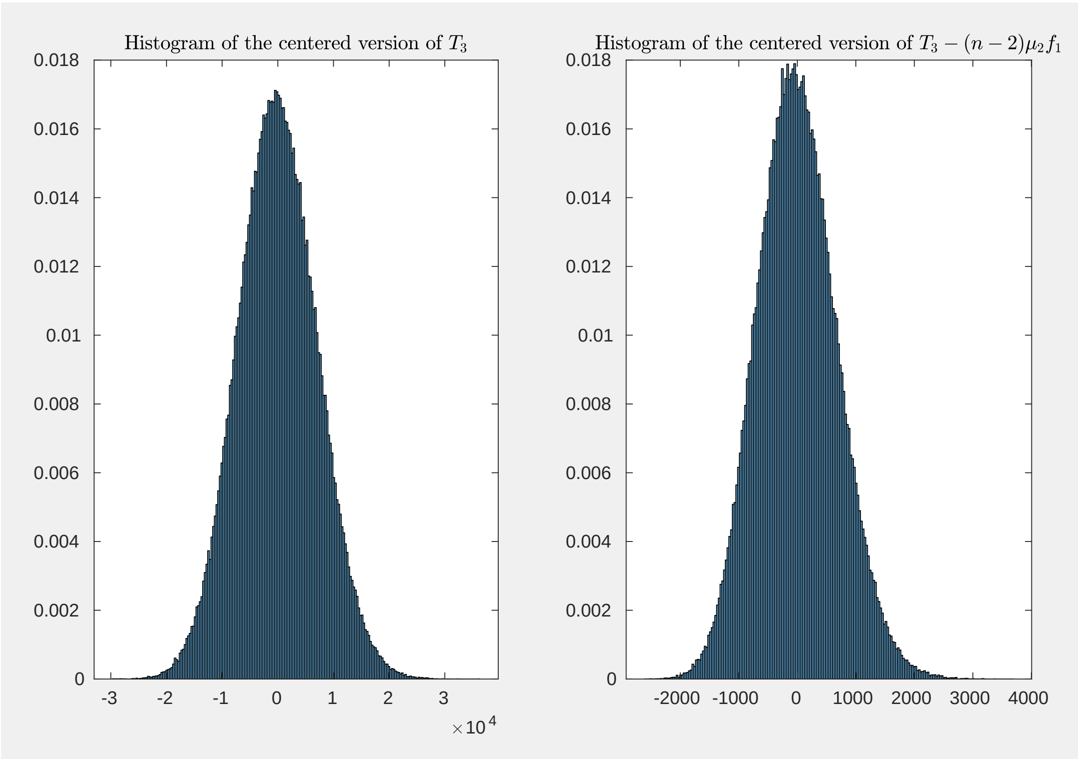

Figure 4.1: The histograms show roughly 700.000 realizations of the ERGM for $n = 100$ and parameters satisfying the conditions of the above theorem. It can be seen that $T_3 - \mathbb{E}_{\mu_\beta} T_3$ takes values on the scale n^2, whereas by subtracting the linear approximation we see values of order $n^{3/2}$.

There is universal constant $C > 0$ such that for any symmetric $n \times n$ matrix $A = (a_{ij})$ and any $t \geq 0$ it holds

$$\mathbb{P}\left(\left|\sum_{i,j} a_{ij} X_i X_j - \operatorname{tr}(A)\right| \geq t\right) \leq 2 \exp\Big(-\frac{1}{C} \min\Big(\frac{t^2}{M^4 \|A\|_{\mathrm{HS}}^2}, \Big(\frac{t}{M^2 \|A\|_{\mathrm{op}}}\Big)^{\frac{\alpha}{2}}\Big)\Big).$$

For the more general result, the sake of clarity we restrict to the case $\alpha = 1$.

Proposition 5.5. *Let $A = (a_{ij})$ be a symmetric $n \times n$ matrix and let $X_1, \ldots, X_n$ be independent, centered random variables which are sub-exponential and $\mathbb{E}\, X_i^2 = 1$. There is a constant $M = M(c, C)$ such that for any $t \geq 0$ it holds*

$$\mathbb{P}\Big(\Big|\sum_{i,j} a_{ij} X_i X_j - \operatorname{tr}(A)\Big| \geq t\Big) \leq 2 \exp\Big(-\frac{1}{C'} \eta(A, t/M^2)\Big),$$

where

$$\eta(A,t) := \min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}, \Big(\frac{t}{\max_{i=1,\ldots,n} \|(a_{ij})_j\|_2}\Big)^{\frac{2}{3}}, \Big(\frac{t}{\|A\|_\infty}\Big)^{\frac{1}{2}}\Big).$$

So, in contrast to sub-Gaussian random variables, the bound for sub-exponential random variables can obtain up to four different regimes. However, in certain situations these reduce to two; for example, consider the identity matrix $A = \mathrm{Id}$, so that $\sum_{i,j} a_{ij} X_i X_j = \sum_{i=1}^n X_i^2$ is a sum of squares of sub-exponential random variables. In this case, we have

$$\mathbb{P}\Big(\big| \sum_{i=1}^n (X_i^2 - \mathbb{E}\, X_i^2) \big| \geq t \Big) \leq 2 \exp\Big(- \frac{1}{C'} \min(t^2/n, t^{1/2}) \Big).$$

The quadratic part t^2/n is an expression of the CLT, whereas the part $t^{1/2}$ comes from large fluctuations of the individual summands.

1.3 Discussion of related literature

In this section, we collect a selection of multilevel concentration inequalities. The first subsection consists of inequalities which can be applied to any function, whereas the second subsection deals with polynomials only.

1.3.1 Concentration inequalities for general functions

Arguably the easiest non-trivial functions of independent random variables are weighted sums, which were studied intensively and are well understood, cf. the paper by Latala [Lat96]. The Hanson–Wright inequality gives concentration inequalities for quadratic forms in independent, sub-Gaussian random variables. However, there has been much interest in probability on Banach spaces in the eighties and nineties (see for example the monograph [LT91]), and so a natural question is to ask for an analogue of the Hanson–Wright inequality on Banach spaces. In the landmark paper [Tal96a], Talagrand provided concentration inequalities for quadratic forms in the Banach space case, i. e. for the random variable

$$Z \coloneqq \Big\| \sum_{i,j=1}^n x_{ij} \varepsilon_i \varepsilon_j \Big\|. \tag{1.9}$$

Here $(B, \|\cdot\|)$ is a Banach space, x_{ij} are elements of B, $x_{ii} = 0$ for all $i = 1, \ldots, n$ and ε_i are independent Rademacher random variables. To state the result, let B_1^* be the unit ball in the dual space B^* with respect to the dual norm

$$\|x^*\|_* \coloneqq \sup_{x \in B : \|x\| \leq 1} x^*(x).$$

The quantities which control the concentration behavior are

$$U \coloneqq \sup_{x^* \in B_1^*} \|(x^*(x_{ij}))_{i,j}\|_{\mathrm{op}} \quad \text{and} \quad V \coloneqq \mathbb{E} \sup_{x^* \in B_1^*} \Big(\sum_{j=1}^n \Big(\sum_{i=1}^n \varepsilon_i x^*(x_{ij}) \Big)^2 \Big)^{1/2}.$$

Theorem 1.3 (Theorem 1.2 in [Tal96a]). *There exists an absolute constant $K > 0$ such that for some median M of Z as in (1.9) and all $t \geq 0$ we have*

$$\mathbb{P}\left(|Z - M| \geq t\right) \leq 2\exp\Big(- \frac{1}{K} \min\Big(\frac{t^2}{V^2}, \frac{t}{U}\Big)\Big). \tag{1.10}$$

The concentration inequality (1.10) can equivalently be considered as a uniform version of (1.8). To see this, use the dual formulation of the norm $\|x\| = \sup_{x^* \in B_1^*} x^*(x)$ to obtain

$$Z = \sup_{x^* \in B_1^*} \sum_{i,j=1}^n \varepsilon_i \varepsilon_j x^*(x_{ij}),$$

which is a supremum of quadratic forms indexed by the set B_1^* (or rather a dense, countable subset of B_1^* to avoid measurability problems).

In contrast to other results of Talagrand, a generalization of Theorem 1.3 causes many technical challenges, and as far as we are aware, there is no proof of the concentration inequality (1.10) using the entropy method, but only control on the deviation of the upper bound. For example, Ledoux proved the following deviation inequality for bounded, independent random variables using the entropy method.

Theorem 1.4 (Theorem 3.1 in [Led97]). *Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent random variables satisfying $|\varepsilon_i| \leq 1$ and let Z be as in (1.9). There is an absolute constant $K > 0$ such that for any $t \geq 0$*

$$\mathbb{P}(Z - \mathbb{E}\, Z \geq t) \leq 2\exp\Big(- \frac{1}{K} \min\Big(\frac{t^2}{\sigma\, \mathbb{E}\, Z + V^2}, \frac{t}{U}\Big)\Big).$$

Here, we set $\sigma := \sup_{\|y\| \leq 1} \sup_{\alpha : |\alpha| \leq 1} \sup_{\beta : |\beta| \leq 1} \sum_{i,j=1}^n \alpha_i \beta_j \langle y, x_{ij} \rangle$.

The quantity σ is unsatisfactory, as it does not appear in Theorem 1.3, and this issue has been resolved later by Boucheron, Lugosi and Massart, see Theorem 1.8.

Another very important contribution of Talagrand was an isoperimetric inequality for the exponential distribution, see [Tal91a, Theorem 1.2]. In [BL97], Bobkov and Ledoux proved a functional form thereof using the entropy method, and generalized it to any probability measure satisfying a Poincaré inequality.

Theorem 1.5 (Corollary 3.2 in [BL97]). *Let μ be a probability measure on a metric space (E,d) (equipped with the Borel σ-algebra) for which a Poincaré inequality*

$$\lambda_1 \mathrm{Var}_\mu(f) \leq \mathbb{E}_\mu |\nabla f|^2$$

for the gradient $|\nabla f|(x) := \limsup_{y \to x} |f(x) - f(y)|/d(x,y)$ holds. There is a constant $K > 0$ depending on λ_1 such that for any bounded function f on E^n satisfying

$$\sum_{i=1}^n |\nabla_i f|^2 \leq \alpha^2 \quad \textit{and} \quad \max_{i=1,\ldots,n} |\nabla_i f| \leq \beta \qquad \mu^n\textit{-almost surely}$$

and any $t \geq 0$ we have

$$\mu^n\big(|f - \mathbb{E}_{\mu^n} f| \geq t\big) \leq 2\exp\Big(-\frac{1}{K}\min\Big(\frac{t^2}{\alpha^2}, \frac{t}{\beta}\Big)\Big).$$

Here, the two-level concentration inequality holds for any (bounded) function f, and the two regimes of its tail decay are controlled by the L^2 and the L^∞ norm of its gradient.

Next, we want to mention the concentration inequalities for functions of independent random variables as developed by Boucheron, Lugosi and Massart [BLM03] as well as by the three authors and Bousquet in [BBLM05]. Let $X_1, \ldots, X_n$ be independent random variables with values in some measurable space $\mathcal{X}$, $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function and $Z := f(X_1, \ldots, X_n)$. Denote by $X_1', \ldots, X_n'$ independent copies of the X_i, define $Z^{(i)} := f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$ and

$$V_+ := \mathbb{E}\Big[\sum_{i=1}^n (Z - Z^{(i)})_+^2 \mid X\Big] \quad \text{and} \quad V_- := \mathbb{E}\Big[\sum_{i=1}^n (Z - Z^{(i)})_-^2 \mid X\Big].$$

Here, x_+ denotes the positive part and x_- the negative part of a real number. With the aid of a general modified logarithmic Sobolev inequality they have established the following theorem.

Theorem 1.6 (Theorem 2 in [BLM03])**.** *For all $\theta > 0$ and $\lambda \in (0, \theta^{-1})$ it holds*

$$\begin{aligned}
\log \mathbb{E}\Big[\exp\big(\lambda(Z - \mathbb{E}\, Z)\big)\Big] &\leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\Big[\exp\Big(\frac{\lambda V_+}{\theta}\Big)\Big], \\
\log \mathbb{E}\Big[\exp\big(-\lambda(Z - \mathbb{E}\, Z)\big)\Big] &\leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\Big[\exp\Big(\frac{\lambda V_-}{\theta}\Big)\Big].
\end{aligned}$$

To get a grasp of what this theorem implies, observe that one can easily recover sub-Gaussian estimate for the upper tail under the condition $V_+ \leq c$ (and for the lower tail under the condition $V_- \leq c$). Furthermore, it also allows to prove two-level concentration if the moment generating function of V_+ can be bounded, which is shown in the following corollary.

Corollary 1.7 (Corollary 4 in [BLM03])**.** *If V_+ is such that for some $a > 0$ and $\lambda \in [0, a^{-1})$ the inequality $\log \mathbb{E} \exp(\lambda V_+) \leq \frac{\lambda}{1 - a\lambda} \mathbb{E}\, V_+$ holds, then for any $t \geq 0$*

$$\mathbb{P}(Z - \mathbb{E}\, Z \geq t) \leq \exp\Big(-\frac{t^2}{4\,\mathbb{E}\, V_+ + \frac{2}{3}(a+1)t}\Big).$$

Theorem 1.6 was at the heart of many of the applications given in [BLM03] (arguably too many to give an account of all of them here), out of which we want to mention the application to the Rademacher chaos (1.9). It gives deviation inequalities for the upper bound similar to Theorem 1.3, but no lower bound of the same form.

Theorem 1.8 (Theorem 17 in [BLM03]). *Let $\mathcal{F}$ be a finite collection of $n \times n$ symmetric matrices with zeroes on its diagonal and the property that $M \in \mathcal{F}$ implies $-M \in \mathcal{F}$ and set*

$$Z := \sup_{M \in \mathcal{F}} \sum_{i,j=1}^{n} M_{ij} \varepsilon_i \varepsilon_j,$$

where ε_i are independent Rademacher random variables. For all $t \geq 0$ it holds

$$\mathbb{P}(Z - \mathbb{E}\, Z \geq t) \leq \exp\Big(- \frac{t^2}{32\, \mathbb{E} \sup_{M \in \mathcal{F}} \sum_i \big(\sum_j M_{ij} \varepsilon_j \big)^2 + \frac{65}{3} \sup_{M \in \mathcal{F}} \|M\|_{\mathrm{op}} t} \Big).$$

All the theorems presented up to now provide second order concentration inequalities in the sense that they deal with functions resembling quadratic forms, and the concentration inequalities are of Hanson–Wright-type. The following theorems contain inequalities that can be used to prove higher order concentration inequalities. The first one is a celebrated result from [BBLM05]. We define the numeric constant $\kappa := \sqrt{e}/(2\sqrt{e} - 2) \approx 1.271$.

Theorem 1.9 (Theorem 2 in [BBLM05]). *Let $X_1, \ldots, X_n$ be independent random variables, f a measurable function and set $Z = f(X_1, \ldots, X_n)$. We have for all $p \geq 2$*

$$\|(Z - \mathbb{E}\, Z)_+\|_p \leq \sqrt{2\kappa p} \|\sqrt{V_+}\|_p,$$
$$\|(Z - \mathbb{E}\, Z)_-\|_p \leq \sqrt{2\kappa p} \|\sqrt{V_-}\|_p.$$

As a sanity check, it is helpful to observe that the condition $V_+ \leq 1$ leads to a sub-Gaussian estimate for the right tail:

$$\mathbb{P}\big(Z - \mathbb{E}\, Z \geq t\big) \leq 2 \exp\Big(- \frac{t^2}{8\kappa e} \Big).$$

Analogously, $V_- \leq 1$ leads to similar estimates for the left tail.

Among other things, Theorem 1.9 implies Rosenthal and Kahane–Khinchine-type inequalities (see Theorems 7 and 8 in [BBLM05]). For the purpose of this exposition, we also want to mention the extension of Theorem 1.3 to arbitrary degrees $d \geq 2$, which requires some definitions. Let $\mathcal{I}_{n,d}$ be the set of all subsets of $[n]$ with d elements, and $\mathcal{T}$ be a set of vectors indexed by $\mathcal{I}_{n,d}$, which is assumed to be compact as a subset of $\mathbb{R}^{\mathcal{I}_{n,d}}$. Define

$$Z_{\mathcal{T}} := Z := \sup_{t \in \mathcal{T}} \Big| \sum_{I \in \mathcal{I}_{n,d}} t_I \prod_{i \in I} X_i \Big|.$$

For $d = 2$, this can also be written as $\sup_{t \in \mathcal{F}} |\sum_{i<j} t_{ij} X_i X_j|$ for some family $\mathcal{F}$ of symmetric $n \times n$ matrices, and is thus a general form of Z as defined in (1.9). Indeed, if $(t_{ij})_{i,j}$ is a symmetric matrix of vectors in B with $t_{ii} = 0$, D is a dense

subset of B_1^* and we set $\mathcal{T} := \{(x^*(t_{ij}))_{i,j} : x^* \in D\}$, this leads to

$$Z_{\mathcal{T}} = \frac{1}{2}\Big\| \sum_{i,j=1}^n t_{ij} X_i X_j \Big\|.$$

For $k \in \{1,\ldots,d\}$ we define the random variables

$$W_k := \sup_{t\in\mathcal{T}} \sup_{\substack{\alpha^1,\ldots,\alpha^k \\ |\alpha^i|\le 1}} \Big| \sum_{J\in\mathcal{I}_{n,d-k}} \prod_{j\in J} X_j \sum_{\substack{i_1,\ldots,i_k: \\ \{i_1,\ldots,i_k\}\cup J\in\mathcal{I}_{n,d}}} \prod_{j=1}^k \alpha^j_{i_j} t_{\{i_1,\ldots,i_k\}\cup J}\Big|,$$

where $|\cdot|$ denotes the Euclidean norm.

Theorem 1.10 (Corollary 4 in [BBLM05])**.** *Let $X_1,\ldots,X_n$ be independent Rademacher random variables. With the same notations as above and $c := \log(2)/(4\kappa) \approx 0.1363$, for any $t \ge 0$ it holds*

$$\mathbb{P}(Z - \mathbb{E}\, Z \ge t) \le 2\exp\Big(- c \min_{k=1,\ldots,d} \Big(\frac{t}{2d\,\mathbb{E}\, W_k}\Big)^{2/k}\Big).$$

Theorem 1.10 provides multilevel deviation inequalities for multilinear functions in Rademacher random variables, where the constants depend on expectations of various “lower-dimensional” multilinear functions. Note that this is a special case of 4.4 if we set $\beta = 0$.

Another important contribution to the theory of higher order concentration inequalities was made by Adamczak and Wolff [AW15]. At the heart of their method are Sobolev-type inequalities of the form

$$\|g(X) - \mathbb{E}\, g(X)\|_p \le L\sqrt{p}\|\nabla g(X)\|_p, \tag{1.11}$$

which (for example) can be derived from a logarithmic Sobolev inequality. In particular, they prove the following concentration result.

Theorem 1.11 (Theorem 1.2 in [AW15])**.** *Assume that a random vector X in $\mathbb{R}^n$ satisfies* (1.11) *for all (appropriate) functions g, and let $f : \mathbb{R}^n \to \mathbb{R}$ be D times differentiable, such that the D-tensor $f^{(D)}$ of all partial derivatives of order D is bounded uniformly. Then for all $t \ge 0$ the inequality*

$$\mathbb{P}(|f(X) - \mathbb{E}\, f(X)| \ge t) \le 2\exp\Big(- \frac{1}{C_D}\eta_f(t)\Big) \tag{1.12}$$

holds for the tail decay function

$$\eta_f(t) = \min\Big(\min_{J\in P_D} \Big(\frac{t}{L^D \sup_x \|f^{(D)}(x)\|_J}\Big)^{\frac{2}{|J|}}, \min_{1\le d\le D-1} \min_{J\in P_d} \Big(\frac{t}{L^d \|\mathbb{E}\, f^{(d)}\|_J}\Big)^{\frac{2}{|J|}}\Big).$$

Here, $f^{(d)}$ is the d-tensor of all partial derivatives, P_d denotes the set of all partitions of $\{1,\ldots,d\}$ and $\|A\|_J$ are certain tensor-product norms. As the definition of these norms is quite involved, we postpone it to Chapter 5.

Albeit very precise, this concentration inequality might be quite difficult to understand. However, one can see that (1.12) is a multilevel concentration inequality, and the coefficients are given by some norms of higher order derivatives of f. It is also possible to simplify this expression by estimating the tensor norms by their Hilbert–Schmidt norms, and this yields a concentration inequality with possibly D different regimes. Inequality (1.12) is optimal in the sense that it can be reversed up to a multiplicative constant for Gaussian measures (see [AW15, Theorem 3]).

1.3.2 Concentration inequalities for polynomials

A special class of functions that has raised considerable attention are polynomials in independent random variables. Among other fields of application, they are of importance in probabilistic combinatorics, where the random variables are $\{0,1\}$-valued and determine whether a certain object is present or absent, and many interesting functions are the number of certain substructures. For example, in the Erdös–Rényi random graph $G(n,p)$ we consider independent Bernoulli random variables $(X_e)_{e \in E(K_n)}$ with expectation p, where $E(K_n)$ is the set of edges in the complete graph on n vertices. A classical question is the number of occurrences of a fixed graph H in the random graph, which itself is a polynomial in the edge variables X_e. As such, the concentration properties of polynomials have been a major topic of research in the last three decades. Here, we give a brief overview of some important results.

We begin by presenting the concentration inequalities by Kim and Vu [KV00] and their extensions by Vu [Vu02]. To do so, we need to introduce some notations. First, it is always possible to represent a polynomial in random variables $X_1, \ldots, X_n$ with maximal power 1 using a weighted hypergraph with vertex set $V = \{1, \ldots, n\}$, edge set $\mathcal{E}$ and weights $(w_e)_{e \in \mathcal{E}}$ as

$$f(X) = \sum_{e \in \mathcal{E}} w_e \prod_{j \in e} X_j. \tag{1.13}$$

For example, $f(X) = X_1 X_3 + 2 X_2 X_4 + 0.1 X_5$ can be represented by $\mathcal{E} = \{\{1,3\}, \{2,4\}, \{5\}\}$ and weights $1, 2$ and 0.1. Clearly, f can be considered as a function on $\mathbb{R}^n$, and for any subset $A \subset \{1, \ldots, n\}$ we let Y_A be the $|A|$-fold partial derivative with respect to all indexes from A, and for $i \in \{0, \ldots, n\}$ we define

$$E_i = \max_{A \subseteq \{1,\ldots,n\} : |A| = i} \mathbb{E}(Y_A).$$

Lastly, set $E = \max_{i=0,\ldots,n} E_i$ and $E' = \max_{i=1,\ldots,n} E_i$.

Theorem 1.12 (Main Theorem in [KV00])**.** *Let $X_1, \ldots, X_n$ be independent random variables with values in $\{0,1\}$, and $f(X)$ be as in (1.13) for some hypergraph H with positive weights. For any $\lambda \geq 1$ it holds*

$$\mathbb{P}\left(|f(X) - \mathbb{E}\, f(X)| > 8^k \sqrt{k!} \sqrt{E E'} \lambda^k\right) \leq C n^{k-1} \exp(-\lambda).$$

Equivalently, for any $\eta \geq 8^k \sqrt{k! E E'}$ *we have*

$$\mathbb{P}\left(|f(X) - \mathbb{E}\, f(X)| \geq \eta\right) \leq C n^{k-1} \exp\Big(- \frac{\eta^{1/k}}{8(EE')^{1/(2k)}}\Big). \tag{1.14}$$

Theorem 1.12 has been enormously successful, and it has been applied in many different situations. Although the definitions might seem intricate, they are quite natural candidates to obtain concentration of measure results, being the maximum *average* effect of a set of coordinates. Actually, it is the presence of an average effect that is appealing in applications, as the maximum effect can be much larger than the average one.

The results have been generalized in [Vu02] (see [Vu02, Theorems 4.2 and 4.11]), but we do not give an account of these results, as they require more definitions and are difficult to describe. From the perspective of multilevel concentration inequalities, (1.14) is quite unnatural, as it provides *one* level of concentration, which also does not correspond to the "natural" choice of order k (as these would produce a decay of $\eta^{2/k}$), but order $2k$. As such, it is unfortunately hardly possible to compare the Kim–Vu inequality to other concentration results for polynomials.

The two papers of Schudy and Sviridenko [SS11; SS12] contain concentration inequalities for polynomials in so-called *moment bounded* random variables $X_1, \ldots, X_n$. Therein, X is called moment bounded with parameter $L > 0$, if for all $i \geq 1$

$$\mathbb{E}|X|^i \leq iL\, \mathbb{E}|X|^{i-1}. \tag{1.15}$$

The multilevel concentration inequalities are expressed in terms of quantities similar to the expectations of the partial derivatives in the Kim–Vu inequalities. More precisely, set for $r \in \{1, \ldots, q\}$

$$\mu_r = \max_{h_0 \subset \{1,\ldots,n\} : |h_0| = r} \mathbb{E}\, \partial_{h_0} f(|X|).$$

Theorem 1.13 (Theorem 1.2 in [SS12])**.** *Let $X_1, \ldots, X_n$ be independent, moment bounded random variables with parameter L, and $f = f(X)$ be a multilinear polynomial as in* (1.13) *with non-negative coefficients $(w_e)_{e \in E}$ and total degree q. For all $t \geq 0$ it holds*

$$\mathbb{P}(|f - \mathbb{E}\, f| \geq t) \leq \exp\Big(2 - \min\Big(\min_{r=1,\ldots,q} \frac{\lambda^2}{\mu_0 \mu_r L^r R^q}, \min_{r=1,\ldots,q} \Big(\frac{\lambda}{\mu_r L^r R^q}\Big)^{1/r}\Big)\Big),$$

where $R \geq 1$ is some absolute constant.

Similar inequalities hold for general polynomials in moment-bounded random variables, one only needs to replace L^r in the denominator by $L^r \Gamma^r$, where Γ is the maximal power of any random variable, and it is also possible to remove the non-negativity assumption of the coefficients.

Theorem 1.13 appears to be stronger than the Kim–Vu inequalities, see [SS12, Section 1.5], and it is applicable for a wider class of random variables. However, it is unclear what kind of random variables satisfy (1.15); in [SS12, Section 7] the authors provide examples of moment-bounded random variables, such as bounded random variables, continuous and log-concave ones, and discrete distributions on $\mathbb{Z}$ satisfying $p_i^2 \geq p_{i-1}p_{i+1}$ (a form of log-concavity for discrete distributions). It is clear that (1.15) implies that Z is sub-exponential, but we are not sure whether this is also a sufficient condition (see also the open question at the end of this thesis).

Thirdly, we want to mention concentration inequalities proven by Adamczak and Wolff in [AW15]. For a random variable X define the sub-Gaussian norm as

$$\|X\|_{\psi_2} := \inf \Big\{ t > 0 : \mathbb{E} \exp(Y^2/t^2) \leq 2 \Big\}.$$

It is easy to see that $\|X\|_{\psi_2} < \infty$ is equivalent to a sub-Gaussian tail decay of the form $\mathbb{P}(|X| \geq t) \leq c \exp(-Ct^2)$ for some constants $c, C > 0$ (see for example [Ver18]). With the same notation of P_d and $\|\cdot\|_J$ as in Theorem 1.11, they prove the following result.

Theorem 1.14 (Theorem 1.4 in [AW15])**.** *Let $X_1, \ldots, X_n$ be independent random variables with $\|X_i\|_{\psi_2} \leq L$ for all $i \in \{1, \ldots, n\}$. For every polynomial $f : \mathbb{R}^n \to \mathbb{R}$ of degree D and $t \geq 0$ we have*

$$\mathbb{P}(|f(X) - \mathbb{E} f(X)| \geq t) \leq 2 \exp\Big(- \frac{1}{C_D} \min_{d=1,\ldots,D} \min_{J \in P_d} \Big(\frac{t}{L^d \|\mathbb{E} f^{(d)}(X)\|_J} \Big)^{\frac{2}{|J|}} \Big),$$

where C_D is a constant depending on D only.

Finally, concentration properties of polynomials have also been investigated for random vectors X with log-concave distributions. Let $X = (X_1, \ldots, X_n)$ be a random vector with values in $\mathbb{R}^n$ and assume that its distribution μ is absolutely continuous with respect to the Lebesgue measure $d\mu(x) = p(x)dx$ and that the logarithm of p is concave. For example, the uniform distribution on any convex set K in $\mathbb{R}^n$ is log-concave. The first result in this regard was proven by Bourgain [Bou91].

Theorem 1.15 (Theorems 1.1 and 1.3 in [Bou91])**.** *For any positive integer d and $p \geq 1$ there is a constant $c(d,p)$ such that if f is any polynomial in n variables of degree d and K is any convex body in $\mathbb{R}^n$ of volume 1, then*

$$\|f\|_p \leq c_{d,p} \|f\|_1.$$

More precisely, there are absolute constants $c, C > 0$ such that

$$\|f\|_{\Psi_{c/d}} \leq C \|f\|_1.$$

Here, $\Psi_{c/d}$ is the Orlicz (quasi-)norm (see Appendix B for details).

This theorem has been generalized by Bobkov [Bob00]. Building upon earlier works of Prokhorov [Pro92; Pro93] and a localization technique developed in [KLS95; LS93], the following extension of Theorem 1.15 was proven.

Theorem 1.16 (Theorems 1 and 2 in [Bob00]). *Let μ be a log-concave probability measure on $\mathbb{R}^n$. There is a constant $C > 0$ such that for any polynomial f of degree d and any $p \geq 2$ it holds*

$$\|f\|_p \leq p^{Cd} \|f\|_1.$$

Furthermore, for some universal constant $C > 0$ we have

$$\|f\|_{\Psi_{1/d}} \leq C^d \|f\|_0 = C^d \exp\Big(\int \log|f| d\mu \Big).$$

Thus far, we have been unable to connect the $\Psi_{1/d}$ norm estimates of Theorem 1.16 with the multilevel concentration inequalities of Theorem 1.14. As Theorem 1.14 can be reversed for the Gaussian measure, these estimates are tight. On the other hand, the class of log-concave measures is clearly more general.

1.3.3 Further topics

The selection of the results given in last two sections is highly subjective, and we are sure that it does not cover all concentration inequalities, or even topics. For example, there are many results on *empirical processes*[1], which we have not mentioned at all. This will be discussed in some detail in Chapter 3.

A second subject that was have not touched upon are concentration properties for U-statistics. Some results prior to the year 2000 can be found in the monograph [PG99], and moment and tail inequalities have been proven in [GLZ00],[Ada06],[AW15]. Since the most precise results require quite involved definitions, we choose not to include these here. In Proposition 4.8 we consider U-statistics in not necessarily independent random variables.

Thirdly, we have not mentioned the topic of the *exact* L^p norm and tail behavior, i. e. two-sided inequalities. For example, there are some results on random variables with log-convex or log-concave tails.[2] The log-convex case is well-understood and was described in [HMO97] for sums of independent random variables and in [KL15] for multilinear forms of higher order. On the other hand, the log-concave case appear to be more difficult to handle, and only partial results are available, see [AL12; GK95; Lat96; Lat99; LŁ03]. Moreover, two-sided estimates for non-negative random variables have been derived in [Mel16] and for chaos of order two in symmetric random variables satisfying the inequality $\|X\|_{2p} \leq A\|X\|_p$ in [Mel19].

[1] An empirical process is a random variables of the type $Z = \sup_{f \in \mathcal{F}} |f(X_1, \ldots, X_n)|$ for a (countable) class $\mathcal{F}$ of measurable functions

[2] A random variable X is said to have log-convex (resp. log-concave) tails if the function $t \mapsto -\log \mathbb{P}(|X| \geq t)$ is convex (resp. concave).

1.4 Outline of the thesis

In Chapter 2, after introducing all the necessary notations, we define difference operators and recall the notion of functional inequalities with respect to a difference operator. We focus our attention on two special difference operators denoted by $\mathfrak{d}$ and $\mathfrak{h}$, which will be used throughout this thesis. Furthermore, we show how logarithmic Sobolev inequalities imply L^p norm estimates, and how these in turn can be translated into concentration inequalities (see Propositions 2.8 and 2.10). Section 2.4 provides almost identical L^p norm estimates for functions of independent random variables, allowing for similar proofs of the concentration inequalities for dependent and independent random variables later. We close the chapter by defining the notion of a weakly dependent spin system and provide concrete examples of such systems, including the Ising and the exponential random graph models.

Afterwards we present Bernstein-type inequalities in Chapter 3. In particular, these results can be interpreted as second order concentration inequalities, i. e. they consist of a sub-Gaussian and a sub-exponential regime. Similar properties hold for so-called self-bounding functions. We also prove a variant of a bounded difference inequality in the special case of multilinear polynomials in independent random variables with values in [0,1] and show some consequences. Finally, we give a proof of Talagrand's concentration inequality for the convex distance on the symmetric group.

Chapter 4 presents concentration inequalities for bounded functions in two different settings: The weakly dependent spin systems, which satisfy a logarithmic Sobolev inequality with respect to $\mathfrak{d}$, and independent random variables. In particular, we obtain multilevel concentration inequalities of any order d, as well as tail estimates for the upper bound for empirical processes in weakly dependent random variables (see Theorem 4.4). In Section 4.2 we provide some applications of the general results, yielding concentration inequalities for U-statistics with a bounded kernel, and for a specific set of polynomials in the Ising model. In the case of exponential random graph models, the general results imply concentration properties of the number of triangles around a first order correction, which in turn can be used to deduce a CLT for the triangle count from the CLT of the number of edges in such models.

Thereafter, in Chapter 5, we shift our focus and consider concentration inequalities for polynomials in independent random variables $X_1, \ldots, X_n$ which have α-sub-exponential tails, i. e. which satisfy $\mathbb{P}(|X_i| \geq t) \leq C_1 \exp(-C_2 t^\alpha)$ for some $\alpha \in (0,1]$. We first present some simplified inequalities for polynomials in such random variables, with an emphasis on quadratic forms $f(X) = \sum_{i,j} a_{ij} X_i X_j$ and multilinear d-forms $f(X) = \sum_{i_1,\ldots,i_d} a_{i_1,\ldots,i_d} X_{i_1} \cdots X_{i_d}$. Section 5.1 contains the refined concentration inequalities, including Hanson–Wright-type inequalities for quadratic forms in Proposition 5.5 and multilevel concentration inequalities in the spirit of Adamczak–Wolff in Theorem 5.6. Again, these results are complemented by several applications in Section 5.2, which extend some known concentration inequalities from the sub-Gaussian to the α-sub-exponential case.

Finally, we formulate a few open questions. Some of these could lead to improvements of our results, or extension to more general settings, whereas others are of more fundamental nature and ask for connections between some of the methods of proving concentration inequalities. For example, an intriguing question is to shed some light on connection between the entropy method and the method of exchangeable pairs as developed by Chatterjee [Cha05]. The former is now well-understood and provides ways of establishing higher order concentration, and the latter one yields explicit constants and can be applied under a Dobrushin uniqueness type condition, but appears to be limited to first order results and Bernstein-type inequalities.

The thesis contains an addendum of three appendices and a bibliography.

CHAPTER 2

Preliminaries

In this chapter, we provide some notations as well as preliminary results that will be used throughout this thesis. This includes the notion of difference operators and logarithmic Sobolev inequalities (LSIs) with respect to these, as well as the approach to translate the LSIs to concentration inequalities. Finally we give the definition of a weakly dependent spin system and provide examples thereof.

2.1 Notations

Measure theoretic notations. We assume that all the random variables under consideration are defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The expectation with respect to $\mathbb{P}$ will be denoted by $\mathbb{E}$. If we work with another probability space $(\mathcal{Y}, \mathcal{F}, \mu)$ we also write $\mathbb{E}_\mu$ for the expectation with respect to μ. The L^p (quasi-)norms for $p \in (0,\infty)$ of a random variable f are given by

$$\|f\|_p := \big(\mathbb{E}|f|^p\big)^{1/p}.$$

However, if it is more convenient, we shall switch to the measure theoretic notation $\int f d\mu = \mathbb{E}_\mu f$.

Usually $\mathcal{Y}$ will denote a product space of the form $\mathcal{Y} = \otimes_{i=1}^n \mathcal{X}_i$ (equipped with the product σ-algebra), and we let $X = (X_1, \ldots, X_n)$ be a $\mathcal{Y}$-valued random vector with distribution μ. To shorten the notation, given any vector $x = (x_j)_{j=1,\ldots,n} \in \mathcal{Y}$, we write $\overline{x}_i = (x_j)_{j \neq i}$, and use the notation $(\overline{x}_i, y_i) \in \mathcal{Y}$ for $\overline{x}_i \in \otimes_{j:j\neq i} \mathcal{X}_j$ and $y_i \in \mathcal{X}_i$.

To focus on the task of showing concentration inequalities, we ignore any measurability issues that may arise. This includes the assumption that all the suprema used in this work are measurable.

Tensors and norms. On the finite dimensional vector space $\mathbb{R}^n$ we define the p-norm for $p \in [1,\infty]$ for any $x = (x_1, \ldots, x_n)$ as

$$|x|_p := \begin{cases} \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p} & p \in [1,\infty), \\ \max_{i=1,\ldots,n} |x_i| & p = \infty. \end{cases}$$

The case $p = 2$ is the *Euclidean norm* and we will simply write $|x|$ instead of $|x|_2$. We denote by $\langle x,y\rangle = \sum_{i=1}^n x_i y_i$ the standard scalar product in $\mathbb{R}^n$. Given any linear operator $A : \mathbb{R}^n \to \mathbb{R}^n$, represented by the matrix $A = (a_{ij})_{i,j=1,\ldots,n}$ we let

$$|A|_{p\to q} \coloneqq \sup_{x\in\mathbb{R}^n : |x|_p = 1} |Ax|_q.$$

The *operator norm* $|A|_{2\to 2}$ will also be denoted by $|A|_{\mathrm{op}}$. Lastly, we define the *Hilbert–Schmidt norm* (also known as *Frobenius norm*)

$$|A|_{\mathrm{HS}} = \Big(\sum_{i,j=1}^n a_{ij}^2\Big)^{1/2}.$$

We frequently deal with d-tensors $A = (a_{i_1\ldots i_d})_{i_1,\ldots,i_d=1,\ldots,n}$ for $d > 2$. A represents the unique multilinear map $A : \mathbb{R}^n \otimes \ldots \otimes \mathbb{R}^n \to \mathbb{R}$ satisfying $A(e_{i_1} \otimes \ldots \otimes e_{i_d}) = a_{i_1\ldots i_d}$, where $\otimes$ denotes the tensor product of vector spaces and $(e_i)_{i=1,\ldots,n}$ is the standard basis of $\mathbb{R}^n$. We define the operator norm of A as

$$|A|_{\mathrm{op}} \coloneqq \sup_{\substack{v^1,\ldots,v^d \\ |v^j|\le 1}} \langle v^1 \cdots v^d, A\rangle = \sup_{\substack{v^1,\ldots,v^d \\ |v^j|\le 1}} \sum_{i_1,\ldots,i_d} v^1_{i_1} \cdots v^d_{i_d} A_{i_1\ldots i_d}, \tag{2.1}$$

using the outer product $(v^1 \cdots v^d)_{i_1\ldots i_d} = \prod_{j=1}^d v^j_{i_j}$. Usually we do not need the interpretation of A as a multilinear map and simply think of A as an array of real numbers. The *(generalized) diagonal* $\Delta = \Delta_d$ comprises all indices which are not pairwise distinct, i. e.

$$\Delta_d \coloneqq \big\{(i_1,\ldots,i_d) \in \{1,\ldots,n\}^d : |\{i_1,\ldots,i_d\}| < d\big\}.$$

If $A_{i_1\ldots i_d} = 0$ for all $(i_1,\ldots,i_d) \in \Delta_d$, we say that A has vanishing diagonal. We call a d-tensor A *symmetric* if for any $\pi \in S_d$ (the perutation group of $\{1,\ldots,d\}$) we have $A_{i_1\ldots i_d} = A_{i_{\pi(1)}\ldots i_{\pi(d)}}$. If A is a random d-tensor we write for any $p \in (0,\infty]$

$$\|A\|_{\mathrm{op},p} = \big(\mathbb{E}|A|_{\mathrm{op}}^p\big)^{1/p} \quad \text{and} \quad \|A\|_{\mathrm{HS},p} = (\mathbb{E}|A|_{\mathrm{HS}}^p)^{1/p}.$$

Miscellaneous For any $d \in \mathbb{N}$, $[d] \coloneqq \{1,\ldots,d\}$ denotes the "integer interval". Given a finite set $\mathcal{I}$ we let $\mathcal{P}(\mathcal{I})$ be the set of all partitions of $\mathcal{I}$ and set $\mathcal{P}_d \coloneqq \mathcal{P}_{[d]}$. Throughout this work, we denote by C an absolute constant and by $C_{l_1,\ldots,l_k}$ a constant that depends on some parameters $l_1,\ldots,l_k$ only. In the proofs, the constants may change from line to line.

We let $x_+ \coloneqq \max(x,0)$ be the positive part of $x \in \mathbb{R}$ and $x_- = \min(x,0)$ be its negative part. For the purpose of brevity we set $x_+^2 \coloneqq (x_+)^2$. Finally, we define $x \wedge y \coloneqq \min(x,y)$.

2.2 Difference operators and functional inequalities

In the introduction we outlined the entropy method, but in the formulation (1.4) we restricted ourselves to probability measures on $\mathbb{R}^n$. The approach can be generalized to an arbitrary metric space (X,d) if we define the *modulus of the gradient* as in Theorem 1.5

$$|\nabla f|(x) = \limsup_{y \to x} \frac{|f(y) - f(x)|}{d(y,x)}, \tag{2.2}$$

and set $|\nabla f|(x) = 0$ for any isolated point $x \in X$. Nonetheless, even this generalization still does not capture all spaces of interest. For example, if we consider discrete sets such as $\mathbb{N}$, any point is isolated and thus a logarithmic Sobolev inequality with a gradient of the form (2.2) cannot hold. However, there have been successful replacements for (1.4) in the discrete setting. In the framework of Markov chains on finite spaces, [DS96] replaced the gradient $|\nabla f|$ by making use of the generator of a (continuous-time) Markov chains. Furthermore, in [BT06] several forms of (modified) logarithmic Sobolev inequalities were investigated.

Continuing these ideas, we work in the framework of *difference operators*, which were introduced in [BG99].

Definition 2.1 (Difference operator)**.** Let $(\mathcal{Y}, \mathcal{A}, \mu)$ be a probability space, and let $\mathcal{F}$ be a subset of $L^\infty(\mu)$ satisfying $af + b \in \mathcal{F}$ for any $f \in \mathcal{F}$ and $a \geq 0, b \in \mathbb{R}$. An operator $\Gamma : \mathcal{F} \to L^\infty(\mu)$ is called a difference operator (on $\mathcal{F}$), if for all $f \in \mathcal{F}$, $a \geq 0$ and $b \in \mathbb{R}$ we have $|\Gamma(af + b)| = a\,|\Gamma(f)|$.

Accordingly, we say that μ satisfies a $\Gamma-\mathrm{LSI}(\sigma^2)$, if for any $f \in \mathcal{F}$ it holds

$$\mathrm{Ent}_\mu(f^2) \leq 2\sigma^2 \int \Gamma(f)^2 d\mu. \tag{2.3}$$

The smallest $\sigma^2 > 0$ such that (2.3) holds is known as the logarithmic Sobolev constant (of μ with respect to Γ).

From the two properties of a difference operator we can infer that a $\Gamma-\mathrm{LSI}(\sigma^2)$ implies a *(Γ-)Poincaré inequality* with constant σ^2. Let $\mathrm{Var}_\mu(f) = \int f^2 d\mu - \left(\int f d\mu\right)^2$ be the variance functional.

Lemma 2.2. *Let μ be a probability measure and Γ difference operator on $\mathcal{F}$ such that $\Gamma-\mathrm{LSI}(\sigma^2)$ holds. For any $f \in \mathcal{F}$ we have the Poincaré-type inequality*

$$\mathrm{Var}_\mu(f) \leq \sigma^2 \int \Gamma(f)^2 d\mu. \tag{2.4}$$

Proof. The proof is very classical and can be deduced from (2.3) by a Taylor expansion of $\Psi(x) := x \log x$, and we follow the proof of [DS96, Lemma 3.1]. Fix an arbitrary $f \in \mathcal{F}$, and observe that due to the properties of $\mathcal{F}$ and Γ, for any $\varepsilon > 0$ we have $f_\varepsilon := 1 + \varepsilon f \in \mathcal{F}$ and $\Gamma(f_\varepsilon)^2 = \varepsilon^2 \Gamma(f)^2$. Furthermore, the Taylor expansion $\Psi(1+x) = x + \frac{x^2}{2} + o(x^2)$ as $x \to 0$ implies

$$\begin{aligned}\mathrm{Ent}_\mu(f_\varepsilon^2) &= \int \Psi(1 + 2\varepsilon f + \varepsilon^2 f^2) d\mu - \Psi\Big(1 + 2\varepsilon \int f d\mu + \varepsilon^2 \int f^2 d\mu\Big) \\ &= 2\varepsilon^2 \int f^2 d\mu - 2\varepsilon^2 \Big(\int f d\mu\Big)^2 + o(\varepsilon^2) \\ &= 2\varepsilon^2 \mathrm{Var}_\mu(f) + o(\varepsilon^2).\end{aligned}$$

Consequently, for any $f \in \mathcal{F}$ and any $\varepsilon > 0$ we have by (2.3)

$$2\varepsilon^2 \mathrm{Var}_\mu(f) + o(\varepsilon^2) \leq 2\sigma^2\varepsilon^2 \int \Gamma(f)^2 d\mu.$$

Dividing both sides by $2\varepsilon^2$ and letting $\varepsilon \to 0$ yields the claim. □

We use a certain L^2-type difference operator $\mathfrak{d}$ to apply the entropy method to weakly dependent random variables. Actually, there is an intimate connection to the framework of Markov chains, which we briefly discuss to illustrate the definition. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be finite spaces, set $\mathcal{Y} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$, and let μ be a probability measure on $\mathcal{Y}$ (not necessarily a product measure, for which many of the concepts can be significantly simplified). To any probability measure on a product space, we can associate a reversible Markov chain $(X^{(k)})_{k \in \mathbb{N}_0}$, which is known as *Glauber dynamics*. It is constructed as follows: Starting at some $x \in \mathcal{Y}$ (i. e. $X^{(0)} = x$), select an index $i \in \{1, \ldots, n\}$ uniformly at random, set $X_j^{(1)} = X_j^{(0)}$ for all $j \neq i$, and sample $X_i^{(1)}$ according to the conditional probability measure $\mu(\cdot \mid \overline{x}_i)$. So, its transition probability is given by

$$P(x,y) = n^{-1} \sum_{i=1}^n \mu(y_i \mid \overline{x}_i) \mathbb{1}_{\forall j \neq i: x_j = y_j},$$

and the *Dirichlet form* (see e. g. [DS96]) is

$$\begin{aligned}\mathcal{E}(f,f) &:= \sum_{x \in \mathcal{Y}} \mu(x) f(x)(f(x) - Pf(x)) = \frac{1}{2} \sum_{x \in \mathcal{Y}} \mu(x) \sum_{y \in \mathcal{Y}} P(x,y)(f(x) - f(y))^2 \\ &= \frac{1}{2n} \sum_{i=1}^n \sum_{x \in \mathcal{Y}} \mu(x) \sum_{y_i \in \mathcal{X}_i} \mu(y_i \mid \overline{x}_i)(f(x) - f(\overline{x}_i, y_i))^2 \\ &=: \frac{1}{n} \int |\mathfrak{d} f|^2 d\mu.\end{aligned}$$

Thus, we have a "natural" difference operator arising from the Glauber dynamics on $\mathcal{Y}$, if we let

$$|\mathfrak{d} f|(x)^2 = \frac{1}{2} \sum_{i=1}^n \int (f(x) - f(\overline{x}_i, y_i))^2 d\mu(y_i \mid \overline{x}_i).$$

We can consider this operator in a broader setting. Let $\mathcal{I}$ be a finite set and let $\mathcal{Y} := \otimes_{i\in\mathcal{I}}\mathcal{X}_i$ be a product of Polish spaces. We recall the disintegration theorem in a special form for product spaces - for the existence we refer to [DM78, Chapter III] and for a modern formulation to [AGS08, Theorem 5.3.1].

Proposition 2.3. *Let $(\mathcal{X}_i)_{i\in\mathcal{I}}$ be Polish spaces and $\mathcal{Y} := \otimes_{i\in\mathcal{I}}\mathcal{X}_i$ its product. Let μ be a Borel probability measure on $\mathcal{Y}$ and $I \subsetneq \mathcal{I}$ arbitrary. There exists a Markov kernel $(\mu(\cdot \mid \overline{x}_I))_{\overline{x}_I\in\overline{\mathcal{Y}}_I}$ such that for any Borel set $A \in \mathcal{B}(\mathcal{Y})$ we have*

$$\mu(A) = \int \mu(A \mid \overline{x}_I) d\overline{\mu}_I(\overline{x}_I).$$

Moreover, the Markov kernel is a family of probability measures on $\mathcal{Y}_I := \bigotimes_{i\in I}\mathcal{X}_i$ and for any $f \in L^1(\mu)$

$$\int f d\mu = \int_{\overline{\mathcal{Y}}_I}\int_{\mathcal{Y}_I} f(\overline{x}_I, y_I) d\mu(y_I \mid \overline{x}_I) d\overline{\mu}_I(\overline{x}_I).$$

This leads to the following definition.

Definition 2.4. Let $(\mathcal{X}_i)_{i\in\mathcal{I}}$ be Polish spaces and μ a measure on $\mathcal{Y} = \otimes_{i\in\mathcal{I}}\mathcal{X}_i$. For any $f \in L^2(\mu)$ let

$$\mathfrak{d}_i f(x) := \Big(\frac{1}{2}\int (f(x) - f(\overline{x}_i, x_i'))^2 d\mu(x_i' \mid \overline{x}_i)\Big)^{1/2},$$
$$\mathfrak{d}_i^+ f(x) := \Big(\frac{1}{2}\int (f(x) - f(\overline{x}_i, x_i'))_+^2\Big)^{1/2}$$

and for any $f \in L^\infty(\mu)$ define

$$\mathfrak{h}_i f(x) := \|f(\overline{x}_i,x_i') - f(\overline{x}_i,x_i'')\|_{L^\infty(\mu(\cdot|\overline{x}_i)\otimes\mu(\cdot|\overline{x}_i))},$$
$$\mathfrak{h}_i^+ f(x) := \sup_{x_i'\in\mathrm{supp}(\mu(\cdot|\overline{x}_i))} (f(x) - f(\overline{x}_i, x_i'))_+.$$

To any $\Gamma \in \{\mathfrak{d}, \mathfrak{d}^+, \mathfrak{h}, \mathfrak{h}^+\}$ we associate the vector $\Gamma f = (\Gamma_i f)_{i\in\mathcal{I}}$ and the difference operator $\Gamma(f) := |\Gamma f| = (\sum_{i\in\mathcal{I}}(\Gamma_i f)^2)^{1/2}$. Using $\mathfrak{h}$, it is possible to define *higher order difference operators* $\mathfrak{h}^{(d)}$ for any $d \in \mathbb{N}$ by iteration. More precisely, by setting

$$\mathfrak{h}_{i_1\dots i_d} f = \mathfrak{h}_{i_1}(\mathfrak{h}_{i_2\dots i_d} f)$$

we obtain a d-tensor $\mathfrak{h}^{(d)} f(x)$ with coordinates $\mathfrak{h}_{i_1\dots i_d} f(x)$, and their Euclidean norm gives rise to a difference operator again. In the next lemmas, we collect some elementary properties.

Lemma 2.5. *For any $f \in L^\infty(\mu)$ and any $d \geq 1$ we have the pointwise estimate*

$$|\mathfrak{h}|\mathfrak{h}^{(d)} f|| \leq |\mathfrak{h}^{(d+1)} f|.$$

Proof. Let $i \in \mathcal{I}$ and $x \in \mathcal{Y}$ be fixed and write $\|\cdot\|_{i,x}$ for the L^∞ norm with respect to $\mu(\cdot \mid \overline{x}_i) \otimes \mu(\cdot \mid \overline{x}_i)$. Using the (reverse) triangle inequality for $|\cdot|$ and $\|\cdot\|_{i,x}$ we obtain

$$\begin{aligned}(\mathfrak{h}_i|\mathfrak{h}^{(d)}f|)^2 &= \Big\| \big|\mathfrak{h}^{(d)}f\big|(\overline{x}_i,y) - \big|\mathfrak{h}^{(d)}f\big|(\overline{x}_i,z)\Big\|_{i,x}^2 \le \Big\| |\mathfrak{h}^{(d)}f(\overline{x}_i,y) - \mathfrak{h}^{(d)}f(\overline{x}_i,z)| \Big\|_{i,x}^2 \\ &= \Big\| \sum_{i_1,\dots,i_d} (\mathfrak{h}_{i_1\dots i_d}f(\overline{x}_i,y) - \mathfrak{h}_{i_1\dots i_d}f(\overline{x}_i,z))^2 \Big\|_{i,x} \le \sum_{i_1,\dots,i_d} (\mathfrak{h}_i\mathfrak{h}_{i_1\dots i_d}f)^2.\end{aligned}$$

Summing over $i \in \mathcal{I}$ and taking the square root yields the result. □

Lemma 2.6. *For any $d \ge 1$ it holds*

$$|\mathfrak{h}^+|\mathfrak{h}^{(d)}f|_{\mathrm{op}}| \le |\mathfrak{h}^{(d+1)}f|_{\mathrm{op}}.$$

Proof. To shorten the notation, let T_i be the formal operator that replaces x_i by x_i', and $T_if(x) = f(\overline{x}_i, x_i')$. For any x, by inserting the vectors $\widetilde{v}^1, \dots, \widetilde{v}^d$ maximizing the supremum (in the operator norm) and using the monotonicity of $x \mapsto x_+$ we obtain

$$\begin{aligned}|\mathfrak{h}^+|\mathfrak{h}^{(d)}f|_{\mathrm{op}}|^2 &= \sum_{i\in\mathcal{I}} \Big\| \big(|\mathfrak{h}^{(d)}f|_{\mathrm{op}} - |\mathfrak{h}^{(d)}T_if|_{\mathrm{op}}\big)_+ \Big\|_{i,\infty}^2 \\ &= \sum_{i\in\mathcal{I}} \Big\| \Big(\sup_{v^j} \langle v^1\cdots v^d, \mathfrak{h}^{(d)}f\rangle - \sup_{v^j}\langle v^1\cdots v^d, \mathfrak{h}^{(d)}T_if\rangle \Big)_+ \Big\|_{i,\infty}^2 \\ &\le \sum_{i\in\mathcal{I}} \Big\| \Big(\langle \widetilde{v}^1\cdots\widetilde{v}^d, \mathfrak{h}^{(d)}f - \mathfrak{h}^{(d)}T_if\rangle \Big)_+ \Big\|_{i,\infty}^2.\end{aligned}$$

The triangle inequality and the dual formulation of the Euclidean norm $|x| = \sup_{y:|y|\le 1}\langle x,y\rangle$ yield

$$\begin{aligned}|\mathfrak{h}^+|\mathfrak{h}^{(d)}f|_{\mathrm{op}}|^2 &\le \sum_{i\in\mathcal{I}} \Big\| \sum_{i_1,\dots,i_d} \widetilde{v}^1_{i_1}\cdots\widetilde{v}^d_{i_d} \Big\|(\mathrm{Id}-T_i)\prod_{j=1}^d(\mathrm{Id}-T_{i_s})f\Big\|_{i_1\cdots i_d,\infty}\Big\|_{i,\infty}^2 \\ &\le \sum_{i\in\mathcal{I}} \Big(\sum_{i_1,\dots,i_d} \widetilde{v}^1_{i_1}\cdots\widetilde{v}^d_{i_d}\mathfrak{h}_{ii_1\cdots i_d}f\Big)^2 \\ &= \Big(\sup_{v^{d+1}:|v^{d+1}|\le 1} \sum_{i_{d+1}\in\mathcal{I}} \sum_{i_1,\dots,i_d} \widetilde{v}^1_{i_1}\cdots\widetilde{v}^d_{i_d}v^{d+1}_{i_{d+1}}\mathfrak{h}_{i_1\cdots i_{d+1}}f\Big)^2 \\ &\le \Big(\sup_{v^1,\dots,v^{d+1}:|v^j|\le 1} \sum_{i_1,\dots,i_{d+1}} v^1_{i_1}\cdots v^{d+1}_{i_{d+1}}\mathfrak{h}_{i_1\cdots i_{d+1}}f\Big)^2 \\ &= |\mathfrak{h}^{(d+1)}f|_{\mathrm{op}}^2.\end{aligned}$$

□

Lemma 2.7. *Assume that μ satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. Then it also satisfies a $\mathfrak{h}^+-\mathrm{LSI}(\sigma^2)$ and a $\mathfrak{h}-\mathrm{LSI}(\sigma^2/2)$.*

Proof. The $\mathfrak{h}^+-\mathrm{LSI}(\sigma^2)$ property follows from

$$\begin{aligned}\int |\mathfrak{d} f|^2 d\mu &= \sum_{i\in\mathcal{I}} \int \iint (f(\overline{x}_i, x_i') - f(\overline{x}_i, x_i''))_+^2 d\mu(x_i' \mid \overline{x}_i) d\mu(x_i'' \mid \overline{x}_i) d\overline{\mu}_i(\overline{x}_i) \\ &\le \sum_{i\in\mathcal{I}} \iint \mathfrak{h}_i^+ f(\overline{x}_i, x_i')^2 d\mu(x_i' \mid \overline{x}_i) d\overline{\mu}_i(\overline{x}_i) \\ &= \sum_{i\in\mathcal{I}} \int (\mathfrak{h}_i^+ f)^2 d\mu = \int |\mathfrak{h}^+ f|^2 d\mu.\end{aligned}$$

To see the second implication, we write

$$\begin{aligned}\int |\mathfrak{d} f|^2 d\mu &= \frac{1}{2}\sum_{i\in\mathcal{I}} \int \iint (f(\overline{x}_i, x_i') - f(\overline{x}_i, x_i''))^2 d\mu(x_i' \mid \overline{x}_i) d\mu(x_i'' \mid \overline{x}_i) d\overline{\mu}_i(\overline{x}_i) \\ &\le \frac{1}{2}\sum_{i\in\mathcal{I}} \int (\mathfrak{h}_i f)^2 d\mu = \frac{1}{2}\int |\mathfrak{h} f|^2 d\mu.\end{aligned}$$

□

2.3 The approach of Aida–Stroock and Adamczak: From functional to concentration inequalities

As mentioned in the introduction there is a second approach to obtaining concentration inequalities from logarithmic Sobolev inequalities proven by Aida and Stroock [AS94]. It allows to control the growth of the L^p norms of a function f in terms of the L^p norms of its gradient $|\nabla f|$, or more generally $\Gamma(f)$ for any operator Γ satisfying a chain rule. The aim of this section is to show how to mimic this in the framework of the difference operator $\mathfrak{d}$ and how to deduce multilevel concentration inequalities.

The first proposition is based on a result by Bobkov, showing in [Bob10, Theorem 2.1] how to adapt the Aida–Stroock argument to “non-local” notions of gradient on graphs.

Proposition 2.8. *Let μ be a measure on a product of Polish spaces satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. For any $f \in L^\infty(\mu)$ and any $p \ge 2$ it holds*

$$\|f\|_p^2 - \|f\|_2^2 \le 2\sigma^2(p-2)\|\mathfrak{d} f\|_p^2, \tag{2.5}$$
$$\|f\|_p^2 - \|f\|_2^2 \le \sigma^2(p-2)\|\mathfrak{h}^+|f|\|_p^2 \le \sigma^2(p-2)\|\mathfrak{h} f\|_p^2. \tag{2.6}$$

Consequently, for any bounded function f we have

$$\|f - \mathbb{E} f\|_p \le (2\sigma^2(p-3/2))^{1/2}\|\mathfrak{d} f\|_p \le (\sigma^2(p-3/2))^{1/2}\|\mathfrak{h} f\|_p, \tag{2.7}$$

and for any positive, bounded function f

$$\|(f - \mathbb{E}\, f)_+\|_p \le (\sigma^2(p-1))^{1/2}\|\mathfrak{h}^+ f\|_p. \tag{2.8}$$

Proof. Let $p > 0$, and f be any measurable function on an arbitrary probability space such that $0 < \|f\|_{p+\varepsilon} < \infty$ for some $\varepsilon > 0$. We have the formula (see e. g. [AS94])

$$\frac{d}{dp}\|f\|_p^2 = \frac{2}{p^2}\|f\|_p^{2-p}\,\mathrm{Ent}(|f|^p). \tag{2.9}$$

Moreover, note that for any $i \in \mathcal{I}$ it holds

$$\begin{aligned}\mathbb{E}_\mu(\mathfrak{d}_i f)^2 &= \frac{1}{2}\iint (f(x) - f(\overline{x}_i, y_i))^2 d\mu(y_i \mid \overline{x}_i) d\mu(x) = \int \mathrm{Var}_{\mu(\cdot|\overline{x}_i)}(f(\overline{x}_i, \cdot)) d\overline{\mu}_i(\overline{x}_i)\\ &= \iiint (f(\overline{x}_i, y_i) - f(\overline{x}_i, z_i))_+^2 d\mu(z_i \mid \overline{x}_i) d\mu(y_i \mid \overline{x}_i) d\overline{\mu}_i(\overline{x}_i)\\ &= \iint (f(x) - f(\overline{x}_i, z_i))_+^2 d\mu(y_i \mid \overline{x}_i) d\mu(x).\end{aligned}$$

Therefore, it follows that

$$\mathbb{E}_\mu|\mathfrak{d} f|^2 = \sum_{i\in\mathcal{I}} \iint (f(x) - f(\overline{x}_i, z_i))_+^2\, d\mu(z_i \mid \overline{x}_i) d\mu(x). \tag{2.10}$$

Now let $p > 2$ and f be non-constant. (The assumption $\|f\|_{p+\varepsilon} < \infty$ is always true since $f \in L^\infty(\mu)$.) Applying the $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ to $g := |f|^{p/2}$ and rewriting this in terms of (2.10) yields

$$\mathrm{Ent}(|f|^p) \le 2\sigma^2 \sum_{i\in\mathcal{I}} \iint (g(x) - g(\overline{x}_i, y_i))_+^2 d\mu(y_i \mid \overline{x}_i) d\mu(x) \tag{2.11}$$

$$= 2\sigma^2 \sum_{i\in\mathcal{I}} \iiint (g(\overline{x}_i, x_i) - g(\overline{x}_i, y_i))_+^2 d\mu(x_i \mid \overline{x}_i) d\mu(y_i \mid \overline{x}_i) d\overline{\mu}_i(\overline{x}_i). \tag{2.12}$$

Using the inequality $(a^{p/2} - b^{p/2})_+^2 \le \frac{p^2}{4} a^{p-2}(a-b)^2$ for all $a,b \ge 0$ and all $p \ge 2$, we obtain

$$(g(x) - g(\overline{x}_i, y_i))_+^2 \le \frac{p^2}{4}(|f|(x) - |f|(\overline{x}_i, y_i))_+^2 |f|^{p-2} \le \frac{p^2}{4}(f(x) - f(\overline{x}_i, y_i))^2 |f|^{p-2}(x),$$

from which it follows in combination with (2.11) that

$$\mathrm{Ent}(|f|^p) \le p^2\sigma^2 \int |f|^{p-2} \sum_{i\in\mathcal{I}} (\mathfrak{d}_i f)^2 d\mu = p^2\sigma^2\, \mathbb{E}_\mu |f|^{p-2} |\mathfrak{d} f|^2$$

and in combination with (2.12) that

$$\mathrm{Ent}(|f|^p) \leq \frac{p^2\sigma^2}{2}\mathbb{E}_\mu |f|^{p-2}|\mathfrak{h}^+|f||^2.$$

Hölder's inequality with exponents $\frac{p}{2}$ and $\frac{p}{p-2}$ applied to the last integral yields

$$\begin{aligned}\mathrm{Ent}(|f|^p) &\leq p^2\sigma^2 \|\mathfrak{d}f\|_p^2 \|f\|_p^{p-2}\\ \mathrm{Ent}(|f|^p) &\leq p^2\frac{p^2\sigma^2}{2}\|\mathfrak{h}^+|f|\|_p^2\|f\|_p^{p-2}.\end{aligned}$$

Plugging this into (2.9), we arrive at the differential inequality $\frac{d}{dp}\|f\|_p^2 \leq 2\sigma^2\|\mathfrak{d}f\|_p^2$ and $\frac{d}{dp}\|f\|_p^2 \leq \sigma^2\|\mathfrak{h}^+|f|\|_p^2$ respectively, which after integration gives (2.5) and (2.6). (2.7) can then be deduced using the Poincaré inequality.

Next, let us prove (2.8). (2.6) shows that for any positive function g we have

$$\|g\|_p^2 \leq \|g\|_2^2 + \sigma^2(p-2)\|\mathfrak{h}^+ g\|_p^2. \tag{2.13}$$

If we set $g = (f - \mathbb{E}_\mu f)_+$, it remains to estimate the right hand side. First off, the Poincaré inequality for $\mathfrak{h}^+$ gives

$$\|g\|_2^2 = \int (f-\mathbb{E}_\mu f)_+^2 d\mu \leq \mathrm{Var}_\mu(f) \leq \sigma^2 \int |\mathfrak{h}^+ f|^2 d\mu \leq \sigma^2 \|\mathfrak{h}^+ f\|_p^2 \tag{2.14}$$

Secondly, the inequality $((a-b)_+ - (c-b)_+)_+ \leq (a-c)_+$ yields

$$\mathfrak{h}_i^+ g(x) = \sup_{x_i'} \big((f(x)-\mathbb{E}_\mu f)_+ - (f(\overline{x}_i, x_i') - \mathbb{E}_\mu f)_+\big)_+ \leq \mathfrak{h}_i^+ f(x),$$

so that

$$\|\mathfrak{h}^+ g\|_p^2 \leq \|\mathfrak{h}^+ f\|_p^2. \tag{2.15}$$

Plugging in (2.14) and (2.15) into (2.13) proves the assertion. □

Remark. Clearly the inequality (2.5) can be proven under the more general assumption that the difference operator Γ satisfies for any $f \in L^\infty(\mu)$ and $p \geq 2$ the inequality

$$\int \Gamma(|f|^{p/2})^2 d\mu \leq cp^2 \int \Gamma(f)^2 |f|^{p-2} d\mu$$

for some $c > 0$. As we are not aware of any other difference operator satisfying this class of inequalities, we chose to present Proposition 2.8 in this form.

There are at least two ways to deduce concentration inequalities of exponential type from growth conditions on the L^p norms of a function f. The first one uses the Taylor series of the exponential function to bound the moment generating function of f, and was proven and used in [GS19] and [BGS19].

Proposition 2.9. *If f is a random variable satisfying for some $\gamma > 0$ and all*

$k \in \mathbb{N}_0$ the inequality

$$\|f\|_k \leq \gamma k, \tag{2.16}$$

then we have

$$\mathbb{E} \exp\Big(\frac{|f|}{2\gamma e}\Big) \leq 2 \tag{2.17}$$

Proof. For any $c > 0$, the Taylor series of the exponential function and (2.16) yield

$$\mathbb{E} \exp(c|f|) = 1 + \sum_{k=1}^{\infty} c^k \frac{\mathbb{E}|f|^k}{k!} \leq 1 + \sum_{k=1}^{\infty} (c\gamma)^k \frac{k^k}{k!} \leq \sum_{k=0}^{\infty} (c\gamma e)^k,$$

where the last inequality follows from $k! \geq (\frac{k}{e})^k$ for all $k \in \mathbb{N}$. Inserting $c = (2\gamma e)^{-1}$ we arrive at (2.17). □

Remark. Actually, using the slightly better estimate $k! \geq (\frac{k}{e})^k \sqrt{2\pi}$ the constant $1/(2\gamma e)$ can be replaced by $\sqrt{2\pi}/((\sqrt{2\pi}+1)\gamma e)$.

The second method was proven in [Ada06; AW15] and applied to many classes of functions (such as U-statistics in independent random variables, functions of random vectors satisfying Sobolev-type inequalities and polynomials in sub-Gaussian random variables), and provides multilevel concentration inequalities (see equation (2.18)). Here, it is stated in the form given in [SS18, Proof of Theorem 3.6] with minor modifications.

Proposition 2.10. *Let I be a finite set, $(\alpha_i)_{i\in I}$ and $(C_i)_{i \in I}$ be a collection of positive real numbers and $s \in [0,2)$. Assume that a random variable f satisfies for all $p \geq 2$ the inequality*

$$\|f\|_p \leq \sum_{i \in I} C_i (p-s)^{\alpha_i}.$$

For all $t \geq 0$ we have

$$\mathbb{P}(|f| \geq t) \leq 2 \exp\Big(- \min\Big(\frac{\log(2)}{2-s}, 1\Big) \min_{i \in I} \Big(\frac{t}{C_k e|I|}\Big)^{\frac{1}{\alpha_i}}\Big). \tag{2.18}$$

Proof. By Chebyshev's inequality we have for any $p \geq 1$

$$\mathbb{P}(|f| \geq e\|f\|_p) \leq \exp(-p). \tag{2.19}$$

Define the function

$$\eta_f(t) := s + \min_{i \in I} \Big(\frac{t}{C_i e|I|}\Big)^{\frac{1}{\alpha_i}},$$

and observe that for all $t > 0$ satisfying $\eta_f(t) \geq 2$ we can estimate

$$e\|f\|_{\eta_f(t)} \leq e \sum_{i\in I} \frac{t}{|I|e} = t,$$

so that an application of equation (2.19) with $p = \eta_f(t)$ yields

$$\mathbb{P}(|f| \geq t) \leq \mathbb{P}(|f| \geq e\|f\|_{\eta_f(t)}) \leq \exp\left(-\eta_f(t)\right).$$

Combining it with the trivial estimate in the case $\eta_f(t) \leq 2$ gives

$$\mathbb{P}(|f| \geq t) \leq \exp(2 - \eta_f(t)).$$

To pass from $\exp(2 - \eta_f(t))$ to $2\exp(-c \min_{i \in I}(\frac{t}{|I|eC_i})^{\frac{1}{\alpha_i}})$ requires a bit of analysis, but the constant $c = \min(\frac{\log(2)}{2-s}, 1)$ can be chosen (in the non-trivial regime $\eta_f(t) \geq 2$). $\square$

2.4 Independent random variables

In the case of independent random variables (or, equivalently, product measures), it is unfortunately not true that these satisfy a $\mathfrak{d}-$LSI with some universal constant, so that both cases can be treated in the same way. Indeed, as can be seen in a slightly more in-depth analysis of the $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ property in Appendix C, it is very restrictive in the sense that it is only true for random variables attaining finitely many values.

Another route to obtain concentration inequalities is to modify the entropy method, which was done in the framework of so-called φ-entropies. The idea to replace the function $\varphi_0(x) := x \log x$ in the definition of the entropy $\mathrm{Ent}_\mu^{\varphi_0}(f) = \mathbb{E}_\mu \varphi_0(f) - \varphi_0(\mathbb{E}_\mu f)$ by other functions φ is present in [Cha04]. In the seminal work [BBLM05] the authors prove inequalities for φ-entropies for power functions $\varphi(x) = |x|^\alpha, \alpha \in (1,2]$, leading to moment inequalities for independent random variables.

Recall that for independent random variables $X_1, \ldots, X_n$ with values in some measurable space $(\mathcal{X}, \mathcal{A})$ and copies $X_1', \ldots, X_n'$ (i. e. $X_1', \ldots, X_n'$ are independent of $X_1, \ldots, X_n$ and X_i' has the same distribution as X_i'), and some measurable function $F : \mathcal{X}^n \to \mathbb{R}$ we set

$$V_+ = V_+(F) := \mathbb{E}\Big(\sum_{i=1}^n (F(X_1, \ldots, X_n) - F(\overline{X}_i, X_i'))_+^2 \mid X\Big) \tag{2.20}$$

and $\kappa := \sqrt{e}/(2\sqrt{e} - 2) \approx 1.2707$.

Theorem 2.11 (Theorem 2 in [BBLM05])**.** *Let $X_1, \ldots, X_n$ be independent random variables, $F = F(X_1, \ldots, X_n)$ a measurable function and V^+ as in* (2.20)*. For any $p \geq 2$ we have*

$$\|(F - \mathbb{E} F)_+\|_p \leq \sqrt{2\kappa p}\|\sqrt{V^+}\|_p.$$

The theorem is proven using an elaborate form of the entropy method, and we choose not to reproduce it. The arguments have been simplified and embedded into the framework of difference operators in [BGS19, Section 2], more precisely [BGS19, Theorem 2.3, Corollary 2.6]. Note that there is a slightly different choice of normalization for $\mathfrak{h}^{\pm}$ in [BGS19] leading to other constants. The next theorem uses the normalization of $\mathfrak{h}$ given in the Section 2.2.

Theorem 2.12 (Corollary 2.6 in [BGS19])**.** *Let $X_1, \ldots, X_n$ be independent random variables and $f = f(X_1, \ldots, X_n) \in L^\infty(\mathbb{P})$. For any $p \geq 2$ it holds*

$$\|(f - \mathbb{E}f)_+\|_p \leq \sqrt{2\kappa p}\, \|\mathfrak{h}^+ f\|_p \quad \textit{and} \quad \|(f - \mathbb{E}f)_-\|_p \leq \sqrt{2\kappa p}\, \|\mathfrak{h}^- f\|_p.$$

Consequently, we have

$$\|f - \mathbb{E}\, f\|_p \leq \sqrt{8\kappa p} \|\mathfrak{h} f\|_p. \tag{2.21}$$

The advantage of Theorem 2.12 is that the difference operators $\mathfrak{h}$ are much easier to iterate. However, there is a price to pay: the function needs to be bounded. Moreover, the difference operator $\mathfrak{h}$ cannot be used to prove statements about functions of supremum-type, i. e. $f(X) = \sup_{t \in \mathcal{T}} g_t(X)$. This issue will be discussed in detail in Chapter 4.

2.5 (Weakly dependent) Spin systems

Spin systems are ubiquitous in the modeling of various phenomena, ranging from toy models to explain ferromagnetism (the Ising and the Potts model, or more generally the random cluster model) to voter models (e. g. interpreting the Ising model as a social choice with binary options and interactions between the agents), various random network models (such as the Erdös-Renyi or the exponential random graph model) and models with *hard constraints* such as the random proper coloring model or the hard-core model.

From the physical viewpoint, a spin system models a collection of particles attaining different states and interacting with each other, so that the complete system consists of a set of *configurations* of the form $\mathcal{X}^{\mathcal{I}}$. Mathematically, a spin system can be described as a probability measure μ on such a product space $\mathcal{Y} := \mathcal{X}^{\mathcal{I}}$, and hard constraints translate into conditions on the support of the probability measure. Here we consider finite spin systems, i. e. the sets $\mathcal{X}$ (the *spins*) and $\mathcal{I}$ (the *sites*) are finite.

Albeit very elementary, these finite spin systems can have a rich dependence structure among the sites. Indeed, many toy models of statistical mechanics are defined as finite spin systems. We are interested in the regimes in which the sites exhibit behavior typical of independent random variables. To this end, we define suitable notions of *weak dependence* which, on the technical side, lead to (modified) logarithmic Sobolev inequalities.

Clearly it is in general not possible to observe strong concentration properties in an arbitrary sequence of probability measures. The trivial example $\mathcal{Y} = \{0,1\}^n$ and $\mu = \frac{1}{2}(\delta_{(0,\ldots,0)} + \delta_{(1,\ldots,1)})$ shows that the linear form $f(x) = \sum_{i=1}^n x_i$ is not concentrated on a $\sqrt{n}$ scale. This, however, cannot be expected, as the concentration of measure phenomenon arises in the independent setting due to the disability of the individual random variables to "align", whereas the spins are perfectly aligned under μ. This leads to the notion of *weakly dependent spin systems*.

Let μ be a spin system on $\mathcal{Y} = \mathcal{X}^{\mathcal{I}}$. Define an *interdependence matrix* $(J_{ij})_{i,j\in\mathcal{I}}$ as any matrix with $J_{ii} = 0$ and such that for any $x, y \in \mathcal{Y}$ with $\overline{x}_j = \overline{y}_j$ we have

$$d_{\mathrm{TV}}(\mu(\cdot \mid \overline{x}_i), \mu(\cdot \mid \overline{y}_i)) \leq J_{ij}. \tag{2.22}$$

Here, d_{TV} is the total variation distance. The matrix J (or any norm thereof) may be interpreted as measuring the *strength of the interactions* between the spins in the spin system μ. In particular, $J = 0$ is an interdependence matrix for product measures μ. Moreover, we need to control the minimal probabilities of the marginal distributions of the spin system μ, and hence define for any subset $S \subsetneq \mathcal{I}$ and $i \notin S$

$$\widetilde{\beta}_{i,S}(\mu) \coloneqq \inf_{\substack{x_S\in\mathcal{X}^S\\ \mu(x_S)>0}} \inf_{\substack{y_{S^c}\in\mathcal{X}^{S^c}\\ \mu(y_{S^c},x_S)>0}} \mu((y_{S^c})_i \mid x_S). \tag{2.23}$$

If $S = \emptyset$, this reads $\widetilde{\beta}_{i,\emptyset}(\mu) = \inf_{y\in\mathcal{Y}:\mu(y)>0} \mu(y_i)$. The interpretation of $\widetilde{\beta}_{i,S}(\mu)$ is straightforward: For any admissible partial configuration $x_S \in \mathcal{X}^S$ all possible marginals are supported on points with probability at least $\widetilde{\beta}_{i,S}(\mu)$. Now let

$$\widetilde{\beta}(\mu) \coloneqq \inf_{S\subsetneq\mathcal{I}} \inf_{i\notin S} \widetilde{\beta}_{i,S}(\mu). \tag{2.24}$$

For example, if μ is a product measure, then $\widetilde{\beta}(\mu)$ is the minimal probability of any atom of the marginals.

Definition 2.13. Let μ be a finite spin system on $\mathcal{Y} = \mathcal{X}^{\mathcal{I}}$. μ is called (α_1, α_2)-*weakly dependent* for some $\alpha_1, \alpha_2 \in (0,1)$, if there exists an interdependence matrix J satisfying

$$\widetilde{\beta}(\mu) \geq \alpha_1 \qquad \text{and} \qquad |J|_{2\to2} \leq 1 - \alpha_2.$$

Remark. If there are no hard constraints, i. e. μ has full support, then $\widetilde{\beta}(\mu)$ can be simplified to

$$\widetilde{\beta}(\mu) = I(\mu) \coloneqq \min_{i\in\mathcal{I}} \min_{y\in\mathcal{Y}} \mu(y_i \mid \overline{y}_i).$$

This can be shown by conditioning for any $S \subset \mathcal{I}$ and any $x_S \in \mathcal{X}^S$

$$\mu(y_i \mid x_S) = \mu(x_S)^{-1} \sum_{z\in\mathcal{X}^{\mathcal{I}\setminus(S\cup i)}} \mu(y_i \mid x_S, z)\mu(x_S, z) \geq I(\mu),$$

and the reverse inequality follows by taking $S = \mathcal{I}\backslash\{j\}$.

The significance of this definition of weak dependence is provided by the following theorem.

Theorem 2.14. *Let μ be an (α_1, α_2)-weakly dependent spin system.*

1. *For any function* $f : \mathcal{Y} \to \mathbb{R}_+$ *vanishing outside of* $\operatorname{supp}\mu = \{x \in \mathcal{Y} : \mu(x) > 0\}$ *we have*

$$\mathrm{Ent}_\mu(f) \leq \frac{1}{\alpha_1 \alpha_2^2} \sum_{i \in \mathcal{I}} \int \mathrm{Ent}_{\mu(\cdot \mid \overline{x}_i)}(f(\overline{x}_i, \cdot)) d\mu(x). \tag{2.25}$$

2. μ *satisfies a* $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ *for* $\sigma^2 \coloneqq \log(\alpha_1^{-1})(\log(2)\alpha_1\alpha_2^2)^{-1}$.

Proof of Theorem 2.14. (1): The entropy tensorization property is proven in Appendix A, see Theorem A.2.

(2): For all $i \in \mathcal{I}$ and $y \in \mathcal{Y}$ with $\mu(y) > 0$, $\mu(\cdot \mid \overline{y}_i)$ is a measure on $\mathcal{X}$ with $\min_{x \in \mathcal{X}} \mu(x \mid \overline{y}_i) \geq \alpha_1$, and so [BT06, Remark 6.6] yields

$$\mathrm{Ent}_{\mu(\cdot \mid \overline{y}_i)}(g^2) \leq 2 \frac{\log(\alpha_1^{-1})}{\log(2)} \mathrm{Var}_{\mu(\cdot \mid \overline{y}_i)}(g),$$

which plugged into equation (2.25) leads to

$$\mathrm{Ent}_\mu(f^2) \leq 2 \frac{\log(\alpha_1^{-1})}{\log(2)\alpha_1\alpha_2^2} \sum_{i \in \mathcal{I}} \int \mathrm{Var}_{\mu(\cdot \mid \overline{y}_i)}(f(\overline{y}_i, \cdot)) d\mu(y) = 2\sigma^2 \int |\mathfrak{d} f|^2 d\mu.$$

□

In the next section, we collect various instances of spin systems that satisfy the conditions of Theorem 2.14.

2.6 Examples of weakly dependent spin systems

The results in Chapter 4 are phrased for arbitrary weakly dependent spin systems. More precisely, we formulate the concentration properties for spin systems which satisfy a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. At the same time, we put a special focus on a number of models for which we establish sufficient conditions for weak dependence and apply our general results. Let us emphasize that any of the models depends on a parameter $n \in \mathbb{N}$, so that we are tacitly considering a sequence of spin systems with a growing number of sites. We will usually suppress this dependence.

It is often easier to define a spin system by its Hamiltonian, i. e. by a function $H : \mathcal{Y} \to \mathbb{R}$. The spin system associated to H is given by the *Gibbs measure*

$$\mu(\sigma) = \mu_H(\sigma) = Z^{-1} \exp(H(\sigma)) \quad \text{for} \quad Z = \sum_{\sigma \in \mathcal{Y}} \exp(H(\sigma)).$$

Note that since $-\infty$ is not in the range of H, Gibbs measures always have full support.

2.6.1 Finite product measures

Maybe the easiest example of a weakly dependent spin system is the setting of independent spins. For finite sets $\mathcal{X}_1, \dots, \mathcal{X}_n$ and probability measures $\mu_1, \dots, \mu_n$, their product $\mu = \otimes_{i=1}^n \mu_i$ is a weakly dependent spin system. We can choose $J = 0$ and $\widetilde{\beta}(\mu) = \min_{i \in [n]} \min_{x_i : \mu_i(x_i) > 0} \mu_i(x_i)$.

2.6.2 The Ising and the Potts model

Our first model - and arguably the most famous one - with dependent spins is the Ising model on n sites, labeled with $[n]$, with *interaction matrix* $J = (J_{ij})_{i,j \in [n]}$ and *external field* $h \in \mathbb{R}^n$. It is the spin system on $\mathcal{Y} := \{-1, +1\}^n$ given by the Hamiltonian

$$H(\sigma) := H_{J,h}(\sigma) = \frac{1}{2}\langle \sigma, J\sigma \rangle + \langle h, \sigma \rangle = \frac{1}{2} \sum_{i,j \in [n]} J_{ij}\sigma_i\sigma_j + \sum_{i \in [n]} h_i \sigma_i.$$

Note that due to $\sigma_i^2 = 1$ the choice of $J_{ii} = 0$ is arbitrary (at least on the level of the measure), and we choose it according to the application we have in mind. Thus, the Ising model with interaction J and external field h is the spin system

$$\mu(\sigma) = \mu_{J,h}(\sigma) = Z^{-1} \exp(H(\sigma)) \quad \text{where} \quad Z = \sum_{\sigma \in \{-1,+1\}^n} \exp(H(\sigma)).$$

For example, the case $J = 0$ gives rise to n independent random variables $\sigma_1, \dots, \sigma_n$ with $\mu(\sigma_i = 1) = \exp(h_i)/(\exp(h_i) + \exp(-h_i))$. On the other hand, the case $h = 0$ and $J_{ij} = \beta$ for some $\beta \in (0,1)$ is the Curie–Weiss model mentioned in Chapter 1.

Proposition 2.15. *If μ denotes the Ising model for (J,h) satisfying $|h|_\infty \le \widetilde{\alpha}$ and $J_{ii} = 0$ and*

$$|J|_{\infty \to \infty} = \max_{i \in [n]} \sum_{j=1}^n |J_{ij}| \le 1 - \alpha, \tag{2.26}$$

then μ is a (α_1, α_2)-weakly dependent spin system, where α_1 and α_2 depend on α and $\widetilde{\alpha}$ only (but not on n).

Condition (2.26) appears in various contexts, especially in the infinite dimensional setting $\mathcal{Y} = \{-1, +1\}^{\mathbb{Z}^d}$ (see for example [Kül03], equations (2.1) and (2.2)). We stick to the common notion and call it the *Dobrushin uniqueness condition.*

Remark. By Theorem 2.14, this can be seen as a generalization of the LSI on $\{-1, +1\}^n$ equipped with the uniform measure, which corresponds to the Ising model without any interactions and external field. Furthermore, for $J = 0$ we obtain n independent random variables $\sigma_1, \dots, \sigma_n$ with $\mathbb{P}(\sigma_i = 1) = \frac{1}{2}(1+\tanh(h_i))$. Thus the logarithmic Sobolev constant necessarily depends on $|h|_\infty$. Indeed, it is known that the LSI constant diverges as $p \to 0$ or $p \to 1$, see e. g. [DS96, Theorem A.1].

In the proof we will need the fact that the conditional probabilities of the Ising model are given by

$$\mu(1 \mid \overline{\sigma}_i) = \frac{1}{2}\Big(1 + \tanh\Big(\sigma_i \sum_j J_{ij}\sigma_j + h_i\sigma_i\Big)\Big). \tag{2.27}$$

Proof of Proposition 2.15. First, let us show that $|J| = (|J|_{ij})_{i,j}$ can be used as an interdependence matrix. Fix $i \neq k$ and $z,y \in \mathcal{Y}$ satisfying $\overline{y}_k = \overline{z}_k$, i. e. $y = T_k z$. Define $\sigma \coloneqq (\overline{z}_i, 1)$ and $m_i(\sigma) \coloneqq \sigma_i \sum_j J_{ij}\sigma_j + h_i\sigma_i$. We have by equation (2.27) and the 1-Lipschitz property of tanh

$$\begin{aligned} d_{TV}\left(\mu(\cdot \mid \overline{z}_i), \mu(\cdot \mid \overline{y}_i)\right) &= \frac{1}{2}|\tanh(m_i(\sigma)) - \tanh(m_i(T_k\sigma))| \\ &\leq \frac{1}{2}|m_i(\sigma) - m_i(T_k\sigma)| = |J_{ik}|. \end{aligned}$$

Now for the interdependence matrix J we have

$$|J|_{2\to 2} \leq \sqrt{|J|_{\infty\to\infty}|J^T|_{\infty\to\infty}} \leq 1 - \alpha,$$

which follows from the general estimate $|\lambda_i(AA^T)| \leq \|AA^T\| \leq \|A\|\|A^T\|$ for any matrix norm $\|\cdot\|$.

The lower bound on the conditional probability follows easily from equation (2.27) and the estimate $\max_i |m_i|_\infty \leq |J|_{\infty\to\infty} + |h|_\infty$. □

A well-known extension of the Ising model (which corresponds to two states -1 and $+1$) to $q > 2$ states is the so-called *Potts model*. For any $q \geq 2$ and $\beta > 0$ consider the probability measure $\mu = \mu_q = \mu_{\beta,q}$ on $[q]^n$ induced by the Hamiltonian

$$H(\sigma) = \frac{\beta}{2n}\sum_{k\in[q]} m_k(\sigma)^2 = \frac{\beta}{2n}\sum_{i\neq j}\delta_{\sigma_i=\sigma_j} + \frac{\beta}{2},$$

where $m_k(\sigma) = \sum_{i\in[n]}\delta_{\sigma_i=k}$ denotes the number of spins with *color* k. The parameter β will be called the (inverse) temperature of the model, so that the Potts model depends on the two parameters q and β.

Proposition 2.16. *Let $q \geq 2$ and $\mu = \mu_{\beta,q}$ be the Potts model with parameters satisfying $2(q-1)e^\beta\beta < 1$. For n large enough, μ is an (α_1, α_2)-weakly dependent system, where α_1, α_2 depend on β and q.*

Proof. Fix two distinct sites $r, s \in [n]$, define $m_{rs,k}(\sigma) \coloneqq \sum_{i\notin\{r,s\}}\delta_{\sigma_i=k}$ and decompose the Hamiltonian as

$$H(\sigma) = \frac{\beta}{2n}\sum_{k\in[q]}(m_{rs,k}(\sigma) + \delta_{\sigma_r=k} + \delta_{\sigma_s=k})^2$$

$$
\begin{aligned}
&= \frac{\beta}{2n}\Big(\sum_{k\in[q]} m_{rs,k}(\sigma)^2 + 2m_{rs,k}(\sigma)(\delta_{\sigma_r=k}+\delta_{\sigma_s=k})\Big) + \frac{\beta}{n} + \frac{\beta}{n}\delta_{\sigma_r=\sigma_s}\\
&= \frac{\beta}{2n}\sum_{k\in[q]} m_{rs,k}(\sigma)^2 + \frac{\beta}{n}\Big(m_{rs,\sigma_r}(\sigma) + m_{rs,\sigma_s}(\sigma) + 1 + \delta_{\sigma_r=\sigma_s}\Big).
\end{aligned}
$$

Note that the first sum does not depend on σ_r and σ_s, and so we can express the conditional probability given $\overline{\sigma_r}$ by

$$
\begin{aligned}
\mu(k \mid \overline{\sigma_r}) &= \frac{\exp\Big(H(\overline{\sigma_r},k)\Big)}{\sum_{\widetilde{k}\in[q]}\exp\Big(H(\overline{\sigma_r},\widetilde{k})\Big)}\\
&= \frac{\exp\Big(\frac{\beta}{2n}\sum_{k'=1}^q m_{rs,k'}(\sigma)^2 + \frac{\beta}{n}m_{rs,k}(\sigma) + \frac{\beta}{n}m_{rs,\sigma_s}(\sigma) + \frac{\beta}{n}\delta_{\sigma_s=k}\Big)}{\sum_{\widetilde{k}\in[q]}\exp\Big(H(\overline{\sigma_r},\widetilde{k})\Big)}\\
&= \frac{1}{1+\sum_{\widetilde{k}:\widetilde{k}\neq k}\exp\Big(\frac{\beta}{n}(m_{rs,\widetilde{k}}(\sigma) - m_{rs,k}(\sigma) + \delta_{\sigma_s=\widetilde{k}} - \delta_{\sigma_s=k})\Big)}\\
&= h\Big(\sum_{\widetilde{k}:\widetilde{k}\neq k}\exp\Big(\frac{\beta}{n}(m_{rs,\widetilde{k}}(\sigma) - m_{rs,k}(\sigma) + \delta_{\sigma_s=\widetilde{k}} - \delta_{\sigma_s=k})\Big)\Big).
\end{aligned}
$$

Here, $h(x) = 1/(1+x)$. In the next step, consider two configurations σ,τ which differ at the site s only, i. e. $\sigma_s \neq \tau_s$ and $\overline{\sigma}_s = \overline{\tau}_s$. Set $\widetilde{\beta} \coloneqq \beta/n$. We have by the 1-Lipschitz property of h

$$
\begin{aligned}
&d_{\mathrm{TV}}(\mu(\cdot \mid \overline{\sigma}_r),\mu(\cdot \mid \overline{\tau}_r)) = \frac{1}{2}\sum_{k\in[q]}|\mu(k \mid \overline{\sigma}_r) - \mu(k \mid \overline{\tau}_r)|\\
&\leq \frac{1}{2}\sum_{k\neq\widetilde{k}}\exp\big(\widetilde{\beta}(m_{rs,\widetilde{k}}(\sigma) - m_{rs,k}(\sigma))\big)\big|\exp\big(\widetilde{\beta}(\delta_{\sigma_s=\widetilde{k}} - \delta_{\sigma_s=k})\big) - \exp\big(\widetilde{\beta}(\delta_{\tau_s=\widetilde{k}} - \delta_{\tau_s=k})\big)\big|\\
&\leq \frac{1}{2}e^{\beta}\sum_{k\neq\widetilde{k}}\big|\exp\big(\widetilde{\beta}(\delta_{\sigma_s=\widetilde{k}} - \delta_{\sigma_s=k})\big) - \exp\big(\widetilde{\beta}(\delta_{\tau_s=\widetilde{k}} - \delta_{\tau_s=k})\big)\big|\\
&=: \frac{1}{2}e^{\beta}\sum_{k\neq\widetilde{k}} I(k,\widetilde{k})\\
&= \frac{1}{2}e^{\beta}\Big(\sum_{\widetilde{k}\neq\sigma_s} I(\sigma_s,\widetilde{k}) + \sum_{\widetilde{k}\neq\tau_s} I(\tau_s,\widetilde{k}) + \sum_{k\neq\sigma_s} I(k,\sigma_s) + \sum_{k\neq\tau_s} I(k,\tau_s)\Big).
\end{aligned}
$$

In the last step, we have used the fact that $I(k,\widetilde{k})$ vanishes whenever $k \notin \{\sigma_s,\tau_s\}$ and $\widetilde{k} \notin \{\sigma_s,\tau_s\}$. By a Taylor expansion, we easily see

$$
\begin{aligned}
I(\sigma_s,\tau_s) = I(\tau_s,\sigma_s) &= 2\widetilde{\beta} + o(\widetilde{\beta}^2)\\
I(\sigma_s,r) = I(\tau_s,r) = I(r,\sigma_s) = I(r,\tau_s) &= \widetilde{\beta} + O(\widetilde{\beta}).
\end{aligned}
$$

Hence we have

$$\begin{aligned} |J|_{\infty\to\infty} &= \max_{r\in[n]} \sum_{s\in[n]} |J_{rs}| \le \frac{1}{2} e^{\beta}(4\widetilde{\beta}(q-1)(n-1) + O_{\beta,q}(n^{-1})) \\ &\le 2(q-1)e^{\beta}\beta + O_{\beta}(n^{-1}), \end{aligned}$$

and for $2(q-1)e^{\beta}\beta < 1$ and large enough n this gives an upper bound on the norm of an interdependence matrix. The lower bound on the conditional probability can be easily obtained from the representation of $\mu(k \mid \overline{\sigma}_r)$ in terms of the function h given above, and we omit the details. □

2.6.3 Random networks: The (vertex-weighted) exponential random graph model

The third spin system we consider is a model of a randomly formed network known as the exponential random graph model. These models have been introduced in [HL81] for directed graphs and further developed in [FS86; Str86]. For a thorough historical overview and asymptotic results we refer to the well-written survey [Cha16] or the lecture notes [Cha17]. The basic idea is to use a weight function on the space of all graphs of size n to increase the probability of certain substructures, such as the number of triangles. As such, the model is able to incorporate dependence between the edges.

We denote by $\mathcal{G}_n$ the set of all simple graphs on n vertices labeled with $[n]$, and set $\mathcal{I}_n := \{(i,j) \in [n]^2 : i < j\}$. For two simple graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ let $N_{G_1}(G_2)$ be the number of graph homomorphisms from G_1 to G_2, i. e. edge-preserving injective maps $\varphi : V_1 \to V_2$.

The exponential random graph model is a parametric family of probability distributions on $\mathcal{G}_n$ for some fixed $n \in \mathbb{N}$. Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_s) \in \mathbb{R}^s$ be a weight vector and $G_1 = (V_1, E_1), \dots, G_s = (V_s, E_s)$ simple, connected graphs. The ERGM with parameters $(\boldsymbol{\beta}, G_1, \dots, G_s)$, denoted by $\mu_{\boldsymbol{\beta}}$, is defined as the spin system on $\{0,1\}^{\mathcal{I}_n}$ associated to the Hamiltonian

$$H_{\boldsymbol{\beta}}(x) := n^2 \sum_{i=1}^{s} \beta_i \frac{N_{G_i}(x)}{n^{|V_i|}}.$$

By convention, we take G_1 to be the complete graph K_2 on two vertices. Note that for $s = 1$ we obtain the Erdös–Rényi model with parameter $p = e^{\beta}(1 + e^{\beta})^{-1}$.

In essence, the Hamiltonian H favors all subgraphs G_i with positive constants β_i, and penalizes graphs that contain many subgraphs G_j with negative β_j. Thus, it produces an exponential tilt towards graphs which are of the prescribed structure, i. e. contain many desirable graphs and few undesirable ones. A special role is assigned to β_1, which decides whether the model favors a large number of edges ($\beta_1 \ge 0$) or not ($\beta_1 < 0$); it can be seen as an external field in the Ising model, with the same sign for every edge.

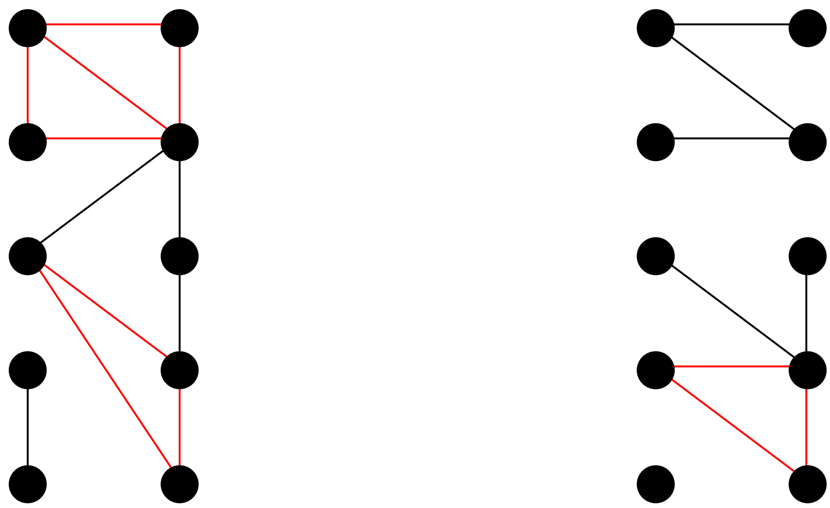

Figure 2.1: Two possible realizations of a random graph on $n = 10$ nodes - the edges that constitute a triangle are colored in red. For example, if $\beta_1 = 0.2, \beta_2 = 0.3$ and G_2 is the triangle graph, then the Hamiltonian is given by $H(x) = 0.4E(x) + 0.18T(x)$, where $E(x)$ denotes the number of edges and $T(x)$ the number of triangles. If we denote by x_1 the left graph and by x_2 the right graph, then $H(x_1) = 5.34$ and $H(x_2) = 3.38$, and so it is roughly seven times more likely to see x_1 than x_2.

For any set of parameters $(\boldsymbol{\beta}, G_1, \ldots, G_s)$ we define the functions $\Phi_{\boldsymbol{\beta}}, \varphi_{\boldsymbol{\beta}} : [0,1] \to \mathbb{R}$ via

$$\Phi_{\boldsymbol{\beta}}(x) = \sum_{i=1}^{s} \beta_i |E_i| x^{|E_i|-1} = \beta_1 + \sum_{i=2}^{s} \beta_i |E_i| x^{|E_i|-1}$$
$$\varphi_{\boldsymbol{\beta}}(x) = \frac{\exp(2\Phi_{\boldsymbol{\beta}}(x))}{1 + \exp(2\Phi_{\boldsymbol{\beta}}(x))} = \frac{1}{2}(1 + \tanh(\Phi_{\boldsymbol{\beta}}(x))),$$

and set $|\boldsymbol{\beta}| \coloneqq (|\beta_1|, \ldots, |\beta_s|)$.

Proposition 2.17. *If $\boldsymbol{\beta}$ is such that $\frac{1}{2}\Phi'_{|\boldsymbol{\beta}|}(1) < 1$, then $\mu_{\boldsymbol{\beta}}$ is a weakly dependent spin system, where α_1 and α_2 depend on $\Phi'_{|\boldsymbol{\beta}|}(1)$.*

To prove the proposition, it is convenient to introduce some notation first. For any graph $x \in \mathcal{G}_n$ and any edge $e = (i,j) \in \mathcal{I}_n$ let x_{e+} (resp. x_{e-}) be the graph with edge set $E(x_{e+}) = E(x) \cup e$ ($E(x_{e-}) = E(x) \backslash e$ respectively). For any function $f : \mathcal{G}_n \to \mathbb{R}$ we define the discrete derivative in the e-th direction as $\partial_e f(x) = f(x_{e+}) - f(x_{e-})$. More generally, given edges $e_1, \ldots, e_k$ we define $\partial_{e_1 \cdots e_k}$

recursively, i. e. $\partial_{e_1\cdots e_k} f = \partial_{e_1}\left(\partial_{e_2\cdots e_k} f\right)$. The partial derivatives of the Hamiltonian are given by

$$\partial_e H(x) = 2\beta_1 + n^2 \sum_{i=2}^{s} \frac{\beta_i}{n^{|V_i|}} (N_{G_i}(x_{e+}) - N_{G_i}(x_{e-})).$$

Now, if $\varphi : G_i \to x_{e_-}$ is a graph homomorphism, then so is $\varphi : G_i \to x_{e_+}$, and so the sum is only nonzero if the edge e is essential for the injection. We write $N_{G_i}(x,e)$ to denote the number of injections of G_i into x which use the edge $e \in E(x)$, so that $\partial_e H(x) = 2\beta_1 + n^2 \sum_{i=2}^{s} \frac{\beta_i}{n^{|V_i|}} N_{G_i}(x,e)$. Especially this gives $|\partial_e H(x)| = O(1)$. Moreover, for two distinct edges $e, f \in \mathcal{I}_n$ we let $N_{G_i}(x,e,f)$ be the number of graph homomorphisms that make use of the edges e and f. We shall require the following lemma, which can be found in [BBS11] and is based on a counting argument.

Lemma 2.18. *With the above notation, for any fixed edge $e \in \mathcal{I}_n$ and simple graph $G_i = (V_i, E_i)$ we have*

$$\sum_{f : f \neq e} N_{G_i}(K_n,f,e) = (|E_i| - 1) N_G(K_n,e) \leq 2|E_i|(|E_i| - 1) n^{|V_i|-2}.$$

Proof of Proposition 2.17. We drop the subscript $\boldsymbol{\beta}$ and set $\mu := \mu_{\boldsymbol{\beta}}$. The lower bound on the conditional probabilities is easy to check, since for any $e \in \mathcal{I}_n$ and any $y \in \mathcal{Y}$

$$\mu(y_e \mid \overline{y}_e) = \frac{1}{2}(1 + \tanh(\partial_e H(y)/2))$$

and $\partial_e H(y) = O(1)$, where the constant depends on $(|\boldsymbol{\beta}|, G_1, \ldots, G_s)$ only. Hence it remains to prove the second condition. To this end, let $x = x_{f+}, y = x_{f-}$ be two graphs which differ in one edge f only, and observe that for each other edge e

$$\begin{aligned} d_{\mathrm{TV}}(\mu(\cdot \mid \overline{x}_e), \mu(\cdot \mid \overline{y}_e)) &= \frac{1}{2}|\tanh(\partial_e H(x_{f+})/2) - \tanh(\partial_e H(x_{f-})/2)| \\ &\leq \frac{1}{4}|\partial_{fe} H(x)| \leq \frac{n^2}{4} \sum_{i=2}^{s} |\beta_i| \frac{N_{G_i}(x,f,e)}{n^{|V_i|}} \\ &\leq \frac{n^2}{4} \sum_{i=2}^{s} |\beta_i| \frac{N_{G_i}(K_n,f,e)}{n^{|V_i|}}, \end{aligned}$$

where K_n is the complete graph on n vertices, and we can choose

$$J_{fe} := \frac{n^2}{4} \sum_{i=2}^{s} |\beta_i| \frac{N_{G_i}(K_n,f,e)}{n^{|V_i|}}.$$

Consequently, after summation in $f \in \mathcal{I}_n$ we obtain by Lemma 2.18

$$\sum_{f : f \neq e} J_{fe} \leq \frac{n^2}{4} \sum_{i=2}^{s} \frac{|\beta_i|}{n^{|V_i|}} \sum_{f \neq e} N_{G_i}(K_n, f, e) \leq \frac{1}{2} \sum_{i=2}^{s} |\beta_i| |E_i| (|E_i| - 1) = \frac{1}{2} \Phi'_{|\beta|}(1).$$

Since the right-hand side is independent of $e \in \mathcal{I}_n$, this yields $|J|_{\infty \to \infty} \leq \frac{1}{2} \Phi'_{|\beta|}(1) < 1$. Moreover, J is a symmetric matrix, so that we have $|J|_{2 \to 2} \leq |J|_{\infty \to \infty}$. □

Another random graph model - the vertex-weighted ERGM - was recently introduced in [DEY19]. The parameter-space is three-dimensional, i. e. $\beta = (\beta_1, \beta_2, p) \in \mathbb{R}^2 \times (0,1)$, and the model is the spin system on $\mathcal{Y} = \{0,1\}^n$ defined via the Hamiltonian

$$H(\sigma) := \log\left(\frac{p}{1-p}\right) \sum_i \sigma_i + \frac{\beta_1}{n} \sum_{i \neq j} \sigma_i \sigma_j + \frac{\beta_2}{n^2} \sum_{i \neq j \neq k} \sigma_i \sigma_j \sigma_k.$$

Note that it resembles the Hamiltonian in the exponential random graph model. On the other hand, it can also be seen as an extension of the Curie–Weiss model on the complete graph with interactions given by a quadratic *and* a cubic form. We define the function

$$\varphi_\beta(x) := \frac{\exp(h_\beta(x))}{1 + \exp_\beta(h(x))} = \frac{\exp\left(\beta_1 x + \beta_2 x^2 + \log(p/(1-p))\right)}{1 + \exp\left(\beta_1 x + \beta_2 x^2 + \log(p/(1-p))\right)}.$$

Similarly to the ERGM, the derivative of φ_β determines whether the system is weakly dependent.

Proposition 2.19. *Let μ_β be the vertex-weighted exponential random graph model and assume that $\sup_{\lambda \in (0,1)} |\varphi'_\beta(\lambda)| < 1$. For n large enough μ_β satisfies the conditions of Theorem 2.14.*

Proof of Proposition 2.19. Since $x_i \in \{0,1\}$ implies $x_i^k = x_i$ for all $k \in \mathbb{N}$, we can rewrite the Hamiltonian using the order parameter $S := \sum_{i=1}^n x_i$ as

$$\mu(x) = Z^{-1} \exp\left(\frac{\beta_1}{n} S(S-1) + \frac{\beta_2}{n^2} S(S-1)(S-2) + \log\frac{p}{1-p} S\right).$$

Observe that we have (with the same notations as in the ERGM)

$$\mu(1 \mid \overline{x}_i) = \frac{\exp(\partial_e H_n(\overline{x}_e, 1))}{1 + \exp(\partial_e H_n(\overline{x}_e, 1))} = \frac{1}{2}\left(1 + \tanh\left(\partial_e H_n(x)/2\right)\right),$$

where in this case $|\partial_e H_n(x)| = |\frac{2\beta_1}{n} \sum_{i \neq e} x_i + \frac{3\beta_2}{n^2} \sum_{i \neq j, i, j \neq e} x_i x_j + \log(p/(1-p))|$ is bounded by a constant depending on β, so that a lower bound on the conditional probabilities holds. The upper bound on the interdependence matrix is already implicitly proven in the proof of [DEY19, Lemma 6], which we modify. Fix a site $e \in \mathcal{I}_n$ and two configurations x, y differing solely at $f \in \mathcal{I}_n$, i. e. $x_f = 1, y_f = 0$, and let $S := \sum_{i=1}^{n} y_i$. We have

$$d_{\mathrm{TV}}(\mu(\cdot \mid \overline{x}_e), \mu(\cdot \mid \overline{y}_e)) = \frac{1}{2}|\tanh(\partial_e H_n(\overline{x}_e, 1)) - \tanh(\partial_e H_n(\overline{y}_e,1))|$$

and since H_n (and as a consequence $\partial_e H_n$) only depends on S, by defining $h_n(\lambda) := \beta_1\lambda + \beta_2\lambda^2 - \frac{\beta_2}{n}\lambda + \log(p/(1-p))$ we can estimate for some $\xi \in (0,1)$

$$J_{fe} \le \left|\frac{\exp(h_n((S+1)/n))}{1+\exp(h_n((S+1)/n))} - \frac{\exp(h_n(S/n))}{1+\exp(h_n(S/n))}\right| = \frac{1}{n}\left|\left(\frac{\exp\circ h_n}{1+\exp\circ h_n}\right)'(\xi)\right|. \tag{2.28}$$

Lastly, if we define $h(\lambda) = \beta_1\lambda + \beta_2\lambda^2 + \log(p/(1-p))$, using the Lipschitz property of the function $\exp(x)/(1+\exp(x))$ it can be shown that

$$\left|\frac{\exp\circ h_n}{1+\exp\circ h_n} - \frac{\exp\circ h}{1+\exp\circ h}\right| = O(n^{-1})$$

and h_n can be replaced by h in (2.28) with an error of $O(n^{-2})$. By summing up over $f \ne e$, we obtain for n large enough and all parameters such that

$$\sup_{\lambda\in(0,1)}\left|\frac{\exp\circ h}{1+\exp\circ h}'\right| < 1 \tag{2.29}$$

that there is an interdependence matrix satisfying $\|J\|_{2\to 2} < 1$. □

Remark. Condition (2.29) can be written in terms of the functions defined for ERGMs. More specifically, we have for any $x \in \mathbb{R}$

$$\frac{\exp(h(x))}{1+\exp(h(x))} = \varphi_{\widetilde{\beta}_1,\widetilde{\beta}_2,\widetilde{\beta}_3}(x)$$

for the ERGM μ_β given by the three parameters $\widetilde{\beta}_1 = \frac{\log(p/(1-p))}{2}$, $\widetilde{\beta}_2 = \frac{\beta_1}{4}$, $\widetilde{\beta}_3 = \frac{\beta_2}{6}$ and the three graphs G_1 an edge, G_2 a 2-star and G_3 a triangle.

2.6.4 Models with exclusion: Random coloring and hard-core model

Given a finite graph $G = (V,E)$ and a set of colors $C = [k]$, the configuration space in the random coloring model is the set of all proper colorings $\Omega_0 \subset C^V$, i. e. the set of all $\sigma \in C^V$ such that $\{v,w\} \in E \Rightarrow \varphi_v \ne \varphi_w$, and $\mu = \mu(G,C)$ denotes the uniform distribution on Ω_0.

Proposition 2.20. *Let $G = (V, E)$ be a simple graph with maximum degree Δ and $k \ge 2\Delta + 1$. Then the conditions of Theorem 2.14 hold for the random coloring model μ_G with α_1, α_2 depending on Δ and k only.*

Proof of Proposition 2.20. Let us first show that $J_{v,w} := \frac{1}{\Delta+1}\mathbb{1}_{v\sim w}$ can be used as an interdependence matrix. To see this, let $c^1, c^2 \in \Omega_0$ be two colorings that differ only in one vertex v_1, and $v_2 \neq v_1$. In the case $v_1 \sim v_2$ the measures $\mu(\cdot \mid \overline{c^i}_{v_2})$ are uniform on $C \backslash \{c^i_{v_k} : v_k \sim v_1\}$ for $i = 1{,}2$, and hence

$$
\begin{aligned}
&d_{\mathrm{TV}}(\mu - G(\cdot \mid \overline{c^1}_{v_2}), \mu_G(\cdot \mid \overline{c^2}_{v_2})) \\
&= \frac{1}{2}\left(\frac{1}{k - |\{\overline{c^1}_{v_2} : v_2 \sim v_1\}|} + \frac{1}{k - |\{\overline{c^2}_{v_2} : v_2 \sim v_1\}|}\right) \leq \frac{1}{k - \Delta} \leq \frac{1}{\Delta + 1}.
\end{aligned}
$$

On the other hand, if $v_2 \not\sim v_1$, then $\mu_G(\cdot \mid \overline{c^i}_{v_2})$ are equal and thus $J_{v_1,v_2} = 0$. So, by the symmetry of J we have

$$
|J|_{2\to 2} \leq |J|_{\infty \to \infty} \leq \max_{v \in V} \sum_{w \in V} J_{v,w} \leq \frac{\Delta}{\Delta + 1} < 1.
$$

Moreover, we have to show $\widetilde{\beta}(\mu_G) \geq c(k,\Delta)$. Let $S \subsetneq V$ be a nonempty collection of vertices, $v_1 \notin S$ and $c^S \in C^S$ be a proper coloring of the induced graph $G\mid_S = (S, E_n \cap S \times S)$ and $c^{v_1} \in C \backslash \{c_{v_2} : v_2 \in S, v_2 \sim v_1\}$. Using the definition $\Omega_0(H)$ for the set of all proper colorings of an arbitrary graph H with a fixed number of colors k, we have

$$
\mu_G(c^{v_1} \mid c^S) = \frac{\mu_G(c^{v_1}, c^S)}{\mu_G(c^S)} = \frac{|\Omega_0(G\mid_S)|}{|\Omega_0(G\mid_{S \cup v_1})|}. \tag{2.30}
$$

It is clear that $|\Omega_0(G\mid_S)| = k^{-1}|\Omega_0(\widetilde{G_S})|$, where $\widetilde{G_S}$ is obtained by adding an isolated vertex v_1 to S. Hence we fix the vertex set $S \cup v_1$ and rewrite equation (2.30) as follows. Let $N(v_1, S) = \{v_2 \in S : v_2 \sim v_1\} = \{e_1, \ldots, e_l\}$ be the neighbors of v_1 in S and for any $w_1, \ldots, w_k \in N(v_1, S)$ we let $e_j = \{v_1, w_j\}$ and $G_{e_1,\ldots,e_k}$ be the graph with edge set $(E_n \cap S \times S) \cup \{e_1, \ldots, e_k\}$. This leads to

$$
\mu_G(c^{v_1} \mid c^S) = k^{-1} \prod_{k=1}^{l} \frac{|\Omega_0(G_{e_1,\ldots,e_{k-1}})|}{|\Omega_0(G_{e_1,\ldots,e_k})|} \geq k^{-1}\left(\frac{\Delta + 1}{\Delta + 2}\right)^l \geq k^{-1}\left(\frac{\Delta + 1}{\Delta + 2}\right)^\Delta,
$$

where the inequality follows from [Jer95, equation (2)], stating that the ratios are bounded from below by a constant $c(\Delta)$.

Finally, the case $S = \emptyset$ is much easier, as $\mu(c_i) = k^{-1}$ due to the invariance of the random coloring model induced by a relabeling of the colors C. □

Another model with hard constraints is the *hard-core model with fugacity parameter* λ. Given a graph $G = (V,E)$ and $\lambda > 0$, the hard-core model $\mu = \mu_{G,\lambda}$ is the spin system on $\mathcal{Y} = \{0{,}1\}^V$ satisfying

$$
\mu(\sigma) = \begin{cases} Z^{-1} \prod_i \lambda^{\sigma_i} & \sigma \text{ admissible} \\ 0 & \text{otherwise} \end{cases}.
$$

Here, an admissible configuration satisfies $\sigma_v \sigma_w = 0$ for all $\{v, w\} \in E$.

Let us first consider the following example, in which the partition function Z

can be easily calculated.

Example 2.21. Consider the star S_n with n neighbors and a hub vertex h. In this case, it is easy to calculate that the normalizing constant is given by

$$Z = \lambda + \sum_{k=0}^{n} \binom{n}{k} \lambda^k = \lambda + (1+\lambda)^n.$$

Here, the λ corresponds to a particle being at the hub and the sum is over all possible configurations of particles at the leaves.

Proposition 2.22. *Let $G = (V,E)$ be any graph with maximum degree Δ and $\lambda < (\Delta - 1)^{-1}$. The conditions of Theorem 2.14 hold for the hard core model μ_λ with fugacity λ, and α_1, α_2 depend on Δ and λ.*

Proof of Proposition 2.22. Since we are going to require hard-core models corresponding to various graphs, we write μ_G to emphasize the graph under consideration. The fugacity λ will not change.

Firstly, let us show that $J_{v_1,v_2} = \frac{\lambda}{1+\lambda}\mathbb{1}_{v_1 \sim v_2}$ can be used as an interdependence matrix. Let $v_1 \in V$ be a site, $\sigma^1, \sigma^2 \in \mathcal{Y}$ be two admissible configurations differing only at v_1, i. e. $\sigma^1_{v_1} = 1, \sigma^2_{v_1} = 0$, and $v_2 \in V$ be another site. If $v_2 \sim v_1$, then $\mu_G(1 \mid \overline{\sigma^1}_{v_1}) = 0$, whereas $\mu_G(1 \mid \overline{\sigma^1}_{v_1}) = \frac{\lambda}{1+\lambda}$. If $v_2 \not\sim v_1$ we have $\mu_G(\cdot \mid \overline{\sigma^1}_{v_1}) = \mu_G(\cdot \mid \overline{\sigma^2}_{v_1})$ and so $J_{v_1,v_2} = 0$. Now, by the symmetry of J the inequality $|J|_{2\to 2} \leq |J|_{\infty\to\infty} \leq \Delta\frac{\lambda}{1+\lambda} < 1$ holds, where the last step is a consequence of $\lambda < \frac{1}{\Delta - 1}$.

Secondly, to see that there is a lower bound on the conditional probabilities, let us first consider the case $S = \emptyset$. Let $v \in V$ be arbitrary, write $N(v)$ for the neighborhood of v and A for the complement of $v \cup N(v)$, and observe that

$$\begin{aligned}\mu_G(\sigma_v = 1) = \mu_G(\sigma_v = 1, \sigma_{N(v)} = 0) &= Z^{-1} \sum_{\widetilde{\sigma}_A} \lambda^{1+|\widetilde{\sigma}_A|} \\ &= \lambda Z^{-1} \sum_{\widetilde{\sigma}_A} \lambda^{|\widetilde{\sigma}_A|} =: \lambda Z^{-1} Z_A,\end{aligned}$$

where the summation is over all admissible (partial) configurations $\widetilde{\sigma}_A$ such that the configuration $(\widetilde{\sigma}_A, 1, 0, \ldots, 0)$ is admissible. Note that due to $\sigma_{N(v)} = 0$ these are actually all admissible configurations of the graph $G\mid_A$ induced by A. The normalization constant Z can be bounded from above and below by

$$(1+\lambda) Z_A \leq Z = \sum_{\widetilde{\sigma}_A \text{ adm.}} \lambda^{|\widetilde{\sigma}_A|} \sum_{\widetilde{\sigma}_{A^c}} \lambda^{|\widetilde{\sigma}_{A^c}|} \mathbb{1}_{(\widetilde{\sigma}_A, \widetilde{\sigma}_{A^c}) \text{ adm.}} \leq (1 + 2^\Delta) Z_A.$$

Here, the first inequality follows by considering the two configurations $\sigma_v = 1, \sigma_{N(v)} = 0$ and $\sigma_{v \cup N(v)} = 0$ only, and the second one is a consequence of the fact that there can be at most $1 + 2^\Delta$ admissible configurations for $v \cup N(v)$. This leads to the bounds

$$\frac{\lambda}{1+2^{\Delta}} \le \mu_G(\sigma_i = 1) \le \frac{\lambda}{1+\lambda}. \tag{2.31}$$

The case $S \neq \emptyset$ follows by a reduction argument. Let $\widetilde{\sigma}_S$ be an admissible configuration of $G\,|_S$, $T := \{w \in S : \sigma_w = 1\} \subset S$ be the occupied sites in S and $N(T) = \cup_{v\in T} N(v)$ be the neighborhood of T. For any $v \in N(T)$ we have $\mu_G(\sigma_v = 1 \mid \widetilde{\sigma}_S) = 0$, so that we only need to consider $v \notin N(T)$.

However, it is an elementary calculation to show that conditional on $\sigma_{N(T)} = 0$, σ_v and σ_T are independent random variables (the easiest way to see this heuristically is to observe that μ only excludes two particles at neighboring vertices, and v and T have distance at least 2). More precisely, we have

$$\mu_G(\sigma_v = 1 \mid \sigma_S = \widetilde{\sigma}_S) = \mu_G(\sigma_v = 1 \mid \sigma_{N(T)} = 0).$$

The probability on the right hand side is equal to the probability that v is occupied in the graph R induced by $[n]\backslash N(T)$, i. e.

$$\mu_G(\sigma_v = 1 \mid \sigma_S = \widetilde{\sigma}_S) = \mu_{G|_R}(\sigma_v = 1).$$

From inequality (2.31) we obtain an upper and lower bound on the probability, leading to $\widetilde{\beta}(\mu_G) \ge c(\Delta, \lambda)$. $\square$

CHAPTER 3

Bernstein-type inequalities

The goal of this chapter is to establish concentration bounds which are akin to Bernstein or Hanson–Wright inequalities. By this, we mean upper bounds of the type

$$\mu\big(f \geq \mathbb{E}_\mu f + t\big) \leq \exp\Big(-\frac{t^2}{2(a+bt)}\Big) \text{ or } \mu\big(f \geq \mathbb{E}_\mu f + t\big) \leq \exp\Big(-\min\Big(\frac{t^2}{a}, \frac{t}{b}\Big)\Big).$$

As both inequalities show two different levels of tail decay (the Gaussian one for $t \leq ab^{-1}$ and an exponential one for $t > ab^{-1}$), we use the terminology of Adamczak [ABW17; AKPS19] and call them *two-level deviation inequalities* (or *concentration* inequalities, if a lower bound holds as well).

Let us first provide some historical remarks on the term Bernstein inequality. As it was not possible to find the original publication of Bernstein [Ber24], we refer to the work of Bennett [Ben62] (however, note that Bennett also did not have access to the publication, but in turn relied on the two works of Craig [Cra33] and Godwin [God55]). In its simplest form (for bounded random variables), it states that if $X_1, \ldots, X_n$ are centered random variables satisfying $|X_i| \leq 1$ for all $i \in [n]$, then for $S_n \coloneqq \sum_{i=1}^n X_i$ and all $t \geq 0$ we have

$$\mathbb{P}(S_n \geq t) \leq \exp\Big(-\frac{t^2}{2\mathrm{Var}(S_n) + \frac{2}{3}Mt}\Big),$$

so that in this case $a = \mathrm{Var}(S_n)$ and $b = \frac{1}{3}M$.

It is striking that Bernstein inequalities appear in many different contexts. Apart from the classical example above, they appear in the study of empirical processes, i. e. random variables of the form

$$Z \coloneqq \sup_{f \in \mathcal{F}} \Big| \sum_{j=1}^n f(X_j) \Big|, \tag{3.1}$$

where X_i are independent random variables and $\mathcal{F}$ is a (countable) class of functions. It was initiated in [Tal96b, Theorem 1.4] using the powerful induction method and convex distance-type inequalities (see [Tal96b, Theorem 4.2]). In subsequent works [Mas00, Theorem 3], [Rio02, Théorème 1.1], [Bou02, Theorem 2.3], [KR05, Theorems 1.1 and 1.2] [Ada08, Theorem 4] either the entropy or the martingale method was applied to prove similar inequalities. For example, it can be deduced from Bousquet's result that for independent and identically distributed random variables $X_1, \ldots, X_n$ and a class of functions $\mathcal{F}$ such that $\mathbb{E} f(X_1) = 0$ for all $f \in \mathcal{F}$ it holds

$$\mathbb{P}(Z - \mathbb{E}\, Z \geq t) \leq \exp\Big(-\nu h\Big(\frac{t}{\nu}\Big)\Big)$$

for $\nu = n \sup_{f \in \mathcal{F}} \operatorname{Var}(f(X_1))$ and $h(x) = (1+x)\log(1+x) - x$. As $h(x) \geq x^2/2(1+x^3)$, this especially implies

$$\mathbb{P}(Z - \mathbb{E}\, Z \geq t) \leq \exp\Big(- \frac{t^2}{2(\nu + \frac{1}{3}t)}\Big),$$

so that the bound holds with $a = \nu$ and $b = \frac{1}{3}$. Note that in this case a stronger estimate, i. e. Bennett's inequality, holds. If one drops the assumption of identical distributions, the result with the best constants was obtained by Klein and Rio, who prove a Bernstein-type inequality for the right tail. We omit the details.

Furthermore, van de Geer and Lederer [GL13] have defined a norm which can precisely capture such Gaussian and exponential behavior, which they termed Bernstein–Orlicz norm. It is defined in terms of the increasing, convex function $\Psi_L(x) := \exp\Big(L^{-1}(\sqrt{1+2Lz} - 1)^2\Big) - 1$ for some fixed $L > 0$, i. e. for any random variable X they set

$$\|X\|_{\Psi_L} = \inf\{t > 0 : \mathbb{E}\, \Psi_L(|X|/t) \leq 1\}.$$

Many of the results given below can be phrased in terms of these Bernstein–Orlicz norms, but for readability's sake we choose not to do so.

Lastly, we want to mention that Bernstein inequalities are not restricted to real value random variables. Oliveira [Oli10] and Tropp [Tro12] independently developed Bernstein-type inequalities for a sum of random matrices, which give concentration inequalities for the largest eigenvalue of a sum of random Hermitian matrices. As we will not need these results, we refer the interested reader to the monograph [Tro15].

3.1 General results

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and let Γ be a difference operator on $\mathcal{A} \subseteq L^\infty(\mu)$. We say that μ satisfies a $\Gamma-$mLSI(ρ), if for all $f \in \mathcal{A}$ we have

$$\operatorname{Ent}_\mu(e^f) \leq \frac{\rho}{2} \mathbb{E}_\mu\, \Gamma(f)^2 e^f. \tag{3.2}$$

We will suppress the dependence of this definition on the class $\mathcal{A}$, as it should be clear from the context.

The first main result are the following Bernstein-type deviation and concentration inequalities.

Theorem 3.1. *Assume that μ satisfies a $\Gamma-\mathrm{mLSI}(\rho)$ for some difference operator Γ and $\rho > 0$. Let f, g be two measurable functions such that $\Gamma(f) \leq g$, and g fulfills $\Gamma(g) \leq b$ for some constant $b > 0$. Then for all $t \geq 0$ the inequality*

$$\begin{aligned} \mu\big(f - \mathbb{E}_\mu f \geq t\big) &\leq \frac{4}{3} \exp\Big(- \frac{1}{8\rho} \min\Big(\frac{t^2}{(\mathbb{E}_\mu g)^2}, \frac{t}{b} \Big)\Big) \\ &\leq \frac{4}{3} \exp\Big(- \frac{t^2}{8\rho((\mathbb{E}_\mu g)^2 + bt)}\Big) \end{aligned} \tag{3.3}$$

holds. If moreover $\Gamma(\lambda f) = |\lambda| \Gamma(f)$ for all $\lambda \in \mathbb{R}$, then for all $t \geq 0$ we have

$$\begin{aligned} \mu\big(|f - \mathbb{E}_\mu f| \geq t\big) &\leq 2 \exp\Big(- \frac{1}{12\rho} \min\Big(\frac{t^2}{(\mathbb{E}_\mu g)^2}, \frac{t}{b} \Big)\Big) \\ &\leq 2 \exp\Big(- \frac{t^2}{12\rho((\mathbb{E}_\mu g)^2 + bt)}\Big). \end{aligned} \tag{3.4}$$

Let us provide some remarks on Theorem 3.1.

Remark. 1) The fact that $\Gamma-\mathrm{mLSI}(\rho)$ provides sub-Gaussian concentration for functions f satisfying $\Gamma(f) \leq 1$ is well known. More precisely, [BG99] shows that in this case

$$\mu(f - \mathbb{E}_\mu f \geq t) \leq \exp\Big(- \frac{t^2}{2\rho}\Big).$$

Furthermore, if Γ is a derivation (i. e. if satisfies $\Gamma(u \circ f) = u'(f)\Gamma(f)$ for any differentiable function u, see also Section 3.2), [BG99] also proves that the exponential moments of λf^2 can be controlled for $t \in [0,1/(2\rho))$, which is expressed in the inequality

$$\int \exp(tf^2) d\mu \leq \exp\Big(\frac{t}{1 - 2ct} \int f^2 d\mu \Big).$$

In contrast, Theorem 3.1 does neither assume Γ to be a derivation, nor that there is a uniform bound on $\Gamma(f)$.

2) The somewhat unsatisfactory factor 4/3 cannot be improved using our method. It is possible to modify our proofs in order to apply [KZ18, Lemma 1.3], which leads to an inequality of the form

$$\mu\big(f - \mathbb{E}_\mu f \geq t\big) \leq \exp\Big(- c \min\Big(\frac{t^2}{\rho(\mathbb{E}_\mu g)^2 + 2b^2\rho^2}, \frac{t}{\sqrt{2}\rho b} \Big)\Big)$$

for some absolute constant c (the same one as in [KZ18]). However, this is at the cost of a weaker denominator in the Gaussian term as compared to (3.3), and so we choose to present it in the form (3.3).

3) It is easy to see that for any $a,b \geq 0$ such that $ab \neq 0$ the function $\varphi_{a,b}(t) = t^2/(2a + 2bt)$ is invertible with inverse $\varphi_{a,b}^{-1}(t) = bt + \sqrt{b^2t^2 + 2at}$, so that the Bernstein-type inequality can equivalently be written as

$$\mu\Big(f - \mathbb{E}_\mu f \geq bt + \sqrt{b^2t^2 + 2at}\Big) \leq \exp(-t) \quad \text{for} \quad t \geq 0.$$

In the same way, $\psi_{a,b}(t) = \min(t^2/a^2, t/b)$ has inverse $\max(\sqrt{a^2t}, bt)$, which allows to rewrite the Hanson–Wright-type inequality in a similar fashion.

Next, we consider a special class of functions, for which analogue results can be obtained - the *self-bounding functions*. In our framework, given a difference operator Γ, we say that $f \geq 0$ is a $(a,b)-$self-bounding function (with respect to Γ) for some $a, b \geq 0$, if

$$\Gamma(f)^2 \leq af + b.$$

For a product measure μ, there are various sources that provide deviation or concentration inequalities for self-bounding functions, see e. g. [BLM00, Theorem 2.1], [Rio01, Théorème 3.1], [BLM03, Theorem 5], [BBLM05, Corollary 1], [Cha05, Theorem 3.9], [MR06, Theorem 1] and [BLM09, Theorem 1]. As many of the proofs rely on the entropy method, it is an easy task to generalize some of the results to the framework of difference operators and obtain Bernstein-type inequalities.

Proposition 3.2. *Assume that μ satisfies a $\Gamma-$mLSI(ρ) and let $f \geq 0$ be a $(a,b)-$self-bounding function. Then we have for all $t \geq 0$*

$$\mu\big(f - \mathbb{E}_\mu f \geq t\big) \leq \exp\Big(-\frac{t^2}{\rho(4a\,\mathbb{E}_\mu f + 4b + \frac{2}{3}at)}\Big).$$

Furthermore, if $\Gamma(\lambda f) = |\lambda|\Gamma(f)$ for all $\lambda \in \mathbb{R}$, then for all $t \in [0, \mathbb{E} f]$ it holds

$$\mu\big(\mathbb{E}_\mu f - f \geq t\big) \leq \exp\Big(-\frac{t^2}{\rho(4a\,\mathbb{E} f + 4b + \frac{2}{3}at)}\Big).$$

As we will show in Proposition 3.10, product measures always satisfy an mLSI with respect to the difference operator used in the works mentioned above. This is a well-known fact and was first proven in [Mas00].

We are also able to prove a version of Talagrand's famous concentration inequality for the convex distance for random permutations by similar means as used in the proofs of the upper results. To this end, recall that for any measurable space Ω and any $\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n$, we may define the convex distance of ω to some measurable set $A \subset \Omega^n$ by

$$d_T(\omega, A) := \sup_{\alpha \in \mathbb{R}^n : |\alpha|_2 = 1} d_\alpha(\omega, A), \quad d_\alpha(\omega, A) := \inf_{\omega' \in A} d_\alpha(\omega, \omega') := \inf_{\omega' \in A} \sum_{i=1}^n |\alpha_i| \mathbb{1}_{\omega_i = \omega_i'}.$$

Proposition 3.3. *Let S_n be the symmetric group and π_n be the uniform distribution on S_n. For any set $A \subseteq S_n$ with $\pi_n(A) \geq 1/2$ and all $t \geq 0$ we have*

$$\pi_n(d_T(\cdot, A) \geq t) \leq 2 \exp\Big(- \frac{t^2}{32} \Big). \tag{3.5}$$

Talagrand's celebrated *convex distance inequality* (see [Tal95, Theorem 5.1]) states that for *any* subset $A \subseteq S_n$ it holds

$$\pi_n(A) \, \mathbb{E}_{\pi_n} \exp \Big(\frac{1}{16} d_T(\cdot, A)^2 \Big) \leq 1, \tag{3.6}$$

which, in particular, easily implies (3.5) with a constant 16 instead of 32. An inequality similar to (3.6) was deduced for product measures in [Tal95], and reproven using the entropy method in [BLM09]. Furthermore, [Pau14] extended Talagrand's inequality to weakly dependent random variables. However, it does not seem possible to adjust the method therein to the case of the symmetric group and so we are not aware of any proof of either of the inequalities using the entropy method. In [Sam17] the author has proven the convex distance inequality for the symmetric group using weak transport inequalities.

Lastly, we show Bernstein-type concentration inequalities for multilinear polynomials in independent random variables with values in [0,1]. We consider a k-homogeneous multilinear form f as follows. Let $H = (V, E, (w_e)_{e \in E})$ be a weighted hypergraph, such that every $e \in E$ consists of exactly k vertices, assume that $(X_v)_{v \in V}$ are independent, [0,1]-valued random variables, and set

$$f(X) = f((X_v)_{v \in V}) = \sum_{e \in E} w_e \prod_{f \in e} X_f = \sum_{e \in E} w_e X_e. \tag{3.7}$$

Define the maximum first order partial derivative $\mathrm{ML}(f)$ as

$$\mathrm{ML}(f) \coloneqq \sup_{v \in V} \sup_{x \in [0,1]^V} \partial_v f(x). \tag{3.8}$$

Proposition 3.4. *Let $(X_v)_{v \in V}$ be independent, $[0,1]$-valued random variables and $f : [0,1]^V \to \mathbb{R}$ given as in (3.7) and assume that $\partial_v f(x) \geq 0$ for all $v \in V$ and $x \in [0,1]^n$. We have for any $t \geq 0$*

$$\mathbb{P}(f(X) - \mathbb{E} f(X) \geq t) \leq \exp \Big(- \frac{t^2}{2k\mathrm{ML}(f)(\mathbb{E} f(X) + t/2)} \Big). \tag{3.9}$$

Furthermore, for $t \in [0, \mathbb{E} f]$ it holds

$$\mathbb{P}(\mathbb{E} f(X) - f(X) \geq t) \leq \exp \Big(- \frac{t^2}{2k\mathrm{ML}(f)} \Big).$$

For example, the monotonicity in every variable condition $\partial_v f(x) \geq 0$ is fulfilled if the weights w_e are non-negative.

The following corollary can be easily deduced from Proposition 3.4

Corollary 3.5. *Let $(X_v)_{v\in V}$ be independent, $[0,1]$-valued random variables and $f = \sum_{l=1}^k f_l$ for some l-homogeneous functions (f_l) of the form (3.7) with positive weights. For any $t \geq 0$ we have*

$$\mathbb{P}(f(X) - \mathbb{E}\, f(X) \geq t) \leq k \exp\Big(- \min_{l=1,\ldots,k} \Big(\frac{t^2}{2l\mathrm{ML}(f_l)(k^2\, \mathbb{E}\, f_l(X) + kt/2)} \Big)\Big).$$

A slight modification of the proof of Proposition 3.4 also allows for deviation inequalities for suprema of such homogeneous polynomials. For example, this can be used to prove concentration inequalities for maxima of linear forms.

Proposition 3.6. *Let $(X_v)_{v\in V}$ be independent, $[0,1]$-valued random variables, $\mathcal{F} \subset \{a \in \mathbb{R}^V : a_i \in [0,1]^n\}$ and define $f(X)_{\mathcal{F}} := \sup_{a\in\mathcal{F}} \sum_{i\in V} a_i X_i$. For any $t \geq 0$ we have*

$$\mathbb{P}(f_{\mathcal{F}}(X) - \mathbb{E}\, f_{\mathcal{F}}(X) \geq t) \leq \exp\Big(- \frac{t^2}{2\sup_{a\in\mathcal{F}} \|a\|_\infty (\mathbb{E}\, f_{\mathcal{F}}(X) + t/2)} \Big).$$

In particular, for any $p \in [1,\infty]$ it holds

$$\mathbb{P}(\|X\|_p - \mathbb{E}\|X\|_p \geq t) \leq \exp\Big(- \frac{t^2}{2(\mathbb{E}\|X\|_p + t/2)} \Big).$$

3.2 Applications

In this section, we describe various situations which give rise to mLSIs with respect to “natural” difference operators, and show some consequences of the main results.

3.2.1 Derivations

If Γ satisfies the chain rule, i. e. for all differentiable $u : \mathbb{R} \to \mathbb{R}$ and $f \in \mathcal{A}$ such that $u \circ f \in \mathcal{A}$ we have $\Gamma(u \circ f) = |u' \circ f|\Gamma(f)$, then (3.2) is equivalent to the usual logarithmic Sobolev inequality

$$\mathrm{Ent}_\mu(f^2) \leq 2\rho\, \mathbb{E}_\mu\, \Gamma(f)^2.$$

Using this, one can derive second order concentration inequalities similar to the ones given in [BCG17] from Theorem 3.1. Let $S^{n-1} := \{x \in \mathbb{R}^n : |x| = 1\}$ be the unit sphere equipped with the uniform measure σ_{n-1}. It is known that for $\rho_n := (n-1)^{-1}$

$$\mathrm{Ent}_{\sigma_{n-1}}(e^f) \leq \frac{\rho_n}{2}\, \mathbb{E}_{\sigma_{n-1}} |\nabla_S f|^2 e^f \tag{3.10}$$

holds for all Lipschitz functions f and the spherical gradient $\nabla_S f$ (see [BCG17, Formula (3.1)] for the logarithmic Sobolev inequality, from which the modified one follows as above). This leads to an alternative proof of [BCG17, Theorem 1.1] with some constant $c = c(b)$.

Proposition 3.7. *Consider S^{n-1} equipped with the uniform measure σ_{n-1} and let $f : S^{n-1} \to \mathbb{R}$ be a C^2 function satisfying $\sup_{\theta \in S^{n-1}} \|f_S''(\theta)\|_{\mathrm{op}} \le 1$. For any $t \ge 0$*

$$\sigma_{n-1}\big(|f - \mathbb{E}_{\sigma_{n-1}} f| \ge t\big) \le 2 \exp\Big(- \frac{1}{12\rho_n} \min\Big(\frac{t^2}{(\mathbb{E}_{\sigma_{n-1}} |\nabla_S f|)^2}, t \Big)\Big). \tag{3.11}$$

In particular, if f is orthogonal to all affine functions, then

$$\sigma_{n-1}\big((n-1)|f - \mathbb{E}_{\sigma_{n-1}} f| \ge t\big) \le 2 \exp\Big(- \frac{1}{12} \min\Big(\frac{t^2}{\mathbb{E}_{\sigma_{n-1}} \|f_S''\|_{\mathrm{HS}}^2}, t \Big)\Big). \tag{3.12}$$

In a similar manner, one may address open subsets of $\mathbb{R}^n$ equipped with some probability measure μ satisfying a logarithmic Sobolev inequality with respect to the usual gradient ∇. This situation has been sketched in [BCG17, Remark 5.3] and was discussed in more detail in [GS20].

Proposition 3.8. *Let $G \subseteq \mathbb{R}^n$ be an open set, equipped with a probability measure μ which satisfies a $\nabla-\mathrm{mLSI}(\rho)$, and let $f : G \to \mathbb{R}$ be a C^2 function satisfying $\sup_{x \in G} \|f''(x)\|_{\mathrm{op}} \le 1$. We have for any $t \ge 0$*

$$\mu\big(|f - \mathbb{E}_\mu f| \ge t\big) \le 2 \exp\Big(- \frac{1}{12\rho} \min\Big(\frac{t^2}{(\mathbb{E}_\mu |\nabla f|)^2}, t \Big)\Big).$$

If we require the first order partial derivatives $\partial_i f$ to be centered (which translates into orthogonality to linear functions if μ is the standard Gaussian measure, for instance), a simple application of the Poincaré inequality yields $\mathbb{E}_\mu |\nabla f|^2 \le \rho\, \mathbb{E}_\mu \|f''\|_{\mathrm{HS}}^2$, which may be used to get back [GS20, Theorem 1.4].

Corollary 3.9. *Let $G \subseteq \mathbb{R}^n$ be an open set, equipped with a probability measure μ satisfying a $\nabla-\mathrm{LSI}(\rho)$, $f : G \to \mathbb{R}$ be a C^2 function and set $g(x) := \langle x - \mathbb{E}_\mu(x), \mathbb{E}_\mu \nabla f \rangle$. If we have*

$$\sup_{x \in \mathrm{supp}(\mu)} \|f''(x)\|_{\mathrm{op}} \le b \quad \textit{and} \quad \int \|f''(x)\|_{\mathrm{HS}}^2 d\mu(x) \le a^2,$$

then for any $t \ge 0$ it holds

$$\mu\big(|f - \mathbb{E}_\mu f - g| \ge t\big) \le 2 \exp\Big(- \frac{1}{12} \min\Big(\frac{t^2}{\rho^2 a^2}, \frac{t}{\rho b} \Big)\Big).$$

Thus, if we recenter a function and its derivatives, the two conditions on the Hessian ensure two-level concentration inequalities. For functions $f(X,Y)$ of independent Gaussian vectors, two-level concentration inequalities have been studied in [Wol13] using the Hoeffding decomposition instead of a centering of the partial derivatives.

3.2.2 Weakly dependent product measures

Next, we show that another functional inequality implies (3.2) with respect to the difference operators $\mathfrak{d}$ and $\mathfrak{d}^+$ (see Definition 2.4). Let μ be a probability measure on a product of Polish spaces $\mathcal{X} = \otimes_{i=1}^n \mathcal{X}_i$ satisfying

$$\mathrm{Ent}_\mu(e^f) \leq \sigma^2 \sum_{i=1}^n \int \mathrm{Cov}_{\mu(\cdot|\overline{x}_i)}(f(\overline{x}_i, \cdot), e^{f(\overline{x}_i,\cdot)}) d\mu(x). \tag{3.13}$$

Here, $\mu(\cdot \mid \overline{x}_i)$ is the regular conditional probability (for the existence see e. g. [AGS08, Theorem 5.3.1]). This functional inequality is (also) known as a modified logarithmic Sobolev inequality in the framework of Markov processes, and it is equivalent to exponential decay of the relative entropy along the Glauber semigroup, see for example [BT06] or [CMT15].

Proposition 3.10. *If μ satisfies (3.13), then a $\mathfrak{d}^+-\mathrm{mLSI}(2\sigma^2)$ and a $\mathfrak{d}-\mathrm{mLSI}(\sigma^2)$ hold. Consequently, for any $f : \mathcal{X} \to \mathbb{R}$ and any $\alpha > \sigma^2$ we have*

$$\mathbb{E}_\mu \exp\big(f - \mathbb{E}_\mu f\big) \leq \big(\mathbb{E}_\mu \exp\big(\alpha|\mathfrak{d}^+ f|^2\big)\big)^{\frac{\sigma^2}{\alpha-\sigma^2}}. \tag{3.14}$$

The same is true for $\mathfrak{d}$ with σ^2 replaced by $\sigma^2/2$. This especially holds for product measures $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ with $\sigma^2 = 1$.

Here, choosing $\alpha = 2\sigma^2$ or $\alpha = \sigma^2$ respectively leads to the exponential inequalities

$$\mathbb{E}_\mu \exp(f) \leq \mathbb{E}_\mu \exp\big(2\sigma^2|\mathfrak{d}^+ f|^2\big) \quad \text{and} \quad \mathbb{E}_\mu \exp(f) \leq \mathbb{E}_\mu \exp\big(\sigma^2|\mathfrak{d} f|^2\big).$$

The first inequality might be considered as a generalization of [Mas00, Lemma 8]. The second inequality is well-known in the case of the discrete cube, cf. [BG99, Corollary 2.4] with a better constant. On the other hand, the proof presented herein is remarkably short and does not rely on some special properties of the measure μ, but can be derived under (3.13).

The property (3.13) is satisfied for a large class containing non-product measures. Note that a sufficient condition (due to Jensen's inequality) for (3.13) is the *approximate tensorization property*

$$\mathrm{Ent}_\mu(e^f) \leq \sigma^2 \sum_{i=1}^n \int \mathrm{Ent}_{\mu(\cdot|\overline{x}_i)}(e^{f(\overline{x}_i,\cdot)}) d\mu(x). \tag{3.15}$$

There are at least two ways of establishing (3.15). The first one is akin to the perturbation argument of Holley and Stroock as outlined in [HS87] (see also [Roy07, Proposition 3.1.18] for a similar reasoning). Assume that $d\mu = Z^{-1} e^f d\nu$, where $f : \mathcal{X} \to \mathbb{R}$ is a measurable function, $\nu = \otimes_{i=1}^n \nu_i$ is *some* product measure and $Z = \mathbb{E}_\nu e^f$. If we require f to be bounded, we clearly have $\mathrm{osc}(f) < \infty$ for its (maximal) oscillation $\mathrm{osc}(f) = \sup_{x\in\mathcal{X}} f(x) - \inf_{x\in\mathcal{X}} f(x)$. Under these assumptions, μ satisfies (3.15) with $\sigma^2 = \exp(2\mathrm{osc}(f))$.

Furthermore, under weak dependence conditions on the local specifications of a measure μ on a product space $\mathcal{Y}$, sufficient conditions to obtain (3.15) were established in [CMT15; Mar13; Mar15].

Note that Proposition 3.10 implies Theorem 1.6, as product measures satisfy (3.13) with $\sigma^2 = 1$. Taking the logarithms on both sides of (3.14) gives for any $\alpha > 1$ and $\lambda \geq 0$

$$\log \mathbb{E}_\mu \exp\Big(\lambda(f - \mathbb{E}_\mu f)\Big) \leq \frac{1}{\alpha - 1} \log \mathbb{E}_\mu \exp\Big(\lambda^2 \alpha |\mathfrak{d}^+ f|^2\Big).$$

It remains to choose some fixed $\theta > 0$ and set $\alpha = (\lambda\theta)^{-1}$.

The next result contains deviation inequalities for suprema of quadratic forms in the spirit of [KZ18] in the weakly dependent case.

Proposition 3.11. *Let μ be supported in $[-1, +1]^n$ and satisfy* (3.13)*. Let $\mathcal{A}$ be a countable class of symmetric matrices, bounded in operator norm and with zeroes on its diagonal. Define $h(x) := \sup_{A \in \mathcal{A}} \langle x, Ax \rangle$, $f_{\mathcal{A}}(x) := \sup_{A \in \mathcal{A}} \|Ax\|$ and $\Sigma := \sup_{A \in \mathcal{A}} \|A\|_{\mathrm{op}}$. We have for any $t \geq 0$*

$$\mu(h - \mathbb{E}_\mu h \geq t) \leq \frac{4}{3} \exp\Big(- \frac{1}{128\sigma^2} \min\Big(\frac{t^2}{2(\mathbb{E}_\mu f_{\mathcal{A}})^2}, \frac{t}{\Sigma} \Big)\Big). \tag{3.16}$$

For a single symmetric matrix A with zeroes on its diagonal and the quadratic form $f(x) = \langle x, Ax \rangle$ similar arguments lead to concentration inequalities for f. The case of a product measure $\mu = \otimes_{i=1}^n \mu_i$ is well-known and we have collected some previous results in the Chapter 1.

Moreover, (3.13) implies an mLSI for convex functions in the spirit of [Led97, Theorem 1.2], and in fact one can recover the results in [Led97] this way with the same proof. A function $f : K \to \mathbb{R}$ on a convex set K is called *separately convex*, if its restriction to one coordinate is convex.

Corollary 3.12. *Assume that μ is supported in $[-1, +1]^n$ and satisfies* (3.13)*. For any differentiable, separately convex function $f : [-1, +1]^n \to \mathbb{R}$ we have*

$$\mathrm{Ent}_\mu(e^f) \leq 4\sigma^2\, \mathbb{E}_\mu |\nabla f|^2 e^f.$$

Especially, if f is separately convex and 1-Lipschitz, this yields

$$\mu(f - \mathbb{E}_\mu f \geq t) \leq \exp\Big(- \frac{t^2}{8\sigma^2}\Big).$$

This can be extended to a class of convex functions with bounded Hessian as follows.

Proposition 3.13. *Let μ be a probability measure supported in $[-1,+1]^n$ satisfying* (3.13) *and f be a convex function such that $\sup_{x \in [-1,+1]^n} |\partial_{ij} f(x)| \leq c_{ij}$. For any*

$t \geq 0$ *it holds*

$$\mu\big(f - \mathbb{E}_\mu f \geq t\big) \leq \frac{4}{3} \exp\Big(- \frac{1}{64\sigma^2} \min\Big(\frac{t^2}{(\mathbb{E}|\nabla f|)^2}, \frac{t}{\|c\|_{\mathrm{op}}} \Big) \Big).$$

3.2.3 Symmetric group

Consider the space S_n of all permutations of $[n]$ equipped with the uniform measure π_n. We write the group operation on S_n as $\tau\sigma$ for $\tau,\sigma \in S_n$. We define two difference operators (on $\mathcal{A} = L^\infty(\pi_n) = \mathbb{R}^{S_n}$)

$$\Gamma(f)(\sigma)^2 = \frac{1}{n} \sum_{i,j=1}^n (f(\sigma) - f(\sigma\tau_{ij}))^2,$$
$$\Gamma_+(f)(\sigma)^2 = \frac{1}{n} \sum_{i,j=1}^n (f(\sigma) - f(\sigma\tau_{ij}))_+^2.$$

We can easily deduce two logarithmic Sobolev inequalities from [GQ03, Theorem 1] as follows. The proof is postponed to Section 3.3.

Proposition 3.14. *Let (S_n, π_n) be the symmetric group equipped with the uniform measure. Then a $\Gamma-$mLSI(1) and a Γ_+-mLSI(2) hold.*

Proposition 3.14 enables us to deduce concentration inequalities for some well-known functions on S_n. Recall that the standard Herbst argument leads to sub-Gaussian concentration results under Lipschitz assumption $\Gamma(f)^2 \leq c^2$, and Proposition 3.2 provides concentration inequalities for self-bounding functions.

Example 3.15. Given a matrix $a = (a_{ij})$ of real numbers satisfying $a_{ij} \in [0,1]$, define $f(\sigma) = \sum_{i=1}^n a_{i,\sigma(i)}$. By elementary computations one can show $\Gamma(f)^2 \leq 4f + 4\,\mathbb{E}_{\pi_n} f$, i. e. f is self-bounding, and furthermore we have $\Gamma(\lambda f) = |\lambda|\Gamma(f)$. As a consequence, Proposition 3.2 leads to the estimate

$$\pi_n\big(|f - \mathbb{E}_{\pi_n} f| \geq t\big) \leq 2\exp\Big(- \frac{t^2}{32\,\mathbb{E}_{\pi_n} f + 8t/3} \Big).$$

Concentration inequalities for f have been proven using the exchangeable pair approach in [Cha05, Proposition 3.10] (see also [Cha07, Theorem 1.1]), with the denominator being $4\,\mathbb{E}_{\pi_n} f + 2t$.

For example, if a is the identity matrix, f is the number of fixed points of a random permutation, which satisfies $\mathbb{E}_{\pi_n} f = 1$ for all $n \in \mathbb{N}$. In this case, f converges to a Poisson distribution with mean 1 as $n \to \infty$ (see e. g. [Dia88]).

In the sequel, we consider concentration properties of Lipschitz functions with respect to some natural metrics on S_n. In particular, we define the four metrics on S_n:

$$H(\pi, \sigma) = \sum_{i=1}^n \mathbb{1}_{\pi(i) \neq \sigma(i)}$$

$$
\begin{aligned}
D(\pi,\sigma) &= \sum_{i=1}^{n} |\pi(i)-\sigma(i)| \\
S(\pi,\sigma) &= \Big(\sum_{i=1}^{n} |\pi(i)-\sigma(i)|^2\Big)^{1/2} \\
I(\pi,\sigma) &= \min\{k \geq 0 : \text{ there are } k \text{ adj. transpositions from } \sigma^{-1} \text{ to } \pi^{-1}\}.
\end{aligned}
$$

The next table collects some basic properties of these functions. We say that a metric is *right invariant*, if for any $\pi,\sigma,\tau \in S_n$ we have $d(\pi,\sigma) = d(\pi\tau,\sigma\tau)$, and *left invariant* if $d(\pi,\sigma) = d(\tau\pi,\tau\sigma)$. It is bi-invariant, if it is right and left invariant.

function d	invariance	mean $\mathbb{E}\, d(\mathrm{id},\cdot)$	$\mathrm{Var}(d(\mathrm{id},\cdot))$	limit theorem
H	bi-invariant	$n-1$	1	$n - H \Rightarrow \mathrm{Poi}(1)$
D	right invariant	$\frac{n^2-1}{3}$	$\frac{(n+1)(2n^2+7)}{45}$	CLT
S^2	right invariant	$\frac{n(n^2-1)}{6}$	$\frac{n^2(n-1)(n+1)^2}{36}$	CLT
I	right invariant	$\frac{n(n-1)}{4}$	$\frac{n(n-1)(2n+5)}{72}$	CLT

Table 3.1: Invariance and probabilistic properties of the four functions D (Spearman's footrule), S^2 (Spearman's rank correlation), H (Hamming distance) and I (Kendall's τ). This table has been extracted from information in [Dia88, Chapter 6].

In the examples below we will make use of the equality

$$\sum_{i \neq j} (\sigma(i) - \sigma(j))^2 = \frac{n^2(n^2-1)}{6},$$

which holds for any permutation $\sigma \in S_n$. We will not give a proof, as it is an easy calculation using $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$ and $\sum_{i=1}^n i = n(n+1)/2$.

Example 3.16. Recall that by the Bobkov–Götze theorem [BG99, Theorem 2.1] (see Theorem 3.25) the $\Gamma-$mLSI(1) implies for any $f : S_n \to \mathbb{R}$, any $\lambda \in \mathbb{R}$ and any $\alpha > 1/2$ the inequality

$$\int \exp\big(\lambda(f - \mathbb{E}_{\pi_n} f)\big) d\pi_n \leq \Big(\int \exp\big(\alpha\lambda^2 \Gamma(f)^2\big) d\pi_n\Big)^{\frac{1}{2\alpha-1}}.$$

We consider *locally Lipschitz functions* f in the sense that for any $\sigma \in S_n$ and any $i,j \in [n]$ we have $|f(\sigma) - f(\sigma\tau_{ij})| \leq d(\sigma,\sigma\tau_{ij})$. In this case, we clearly have $\Gamma(f)^2 \leq n^{-1}\sum_{i,j} d(\sigma,\sigma\tau_{ij})^2$, so that if we define the observable diameter $\mathrm{ObsDiam}(S_n,d) := \max_\sigma n^{-1}\sum_{i,j} d(\sigma,\sigma\tau_{ij})^2$, this in turn yields

$$\int \exp\big(\lambda(f - \mathbb{E}_{\pi_n} f)\big) d\pi_n \leq \exp\big(\lambda^2 \mathrm{ObsDiam}(S_n,d)/2\big).$$

Here, to obtain the constant $1/2$, one has to let $\alpha \to \infty$. Thus, any locally Lipschitz function f satisfies the sub-Gaussian tail estimate

$$\pi_n(|f - \mathbb{E}_{\pi_n} f| \geq t) \leq 2\exp\Big(-\frac{t^2}{2\mathrm{ObsDiam}(S_n,d)}\Big). \tag{3.17}$$

Especially, by integration by parts this also implies a variance estimate

$$\sup_f \mathrm{Var}(f) \leq C\mathrm{ObsDiam}(S_n, d),$$

where the supremum runs over all 1-Lipschitz functions with respect to d.

Furthermore, if the distance d is right invariant, then $\mathrm{ObsDiam}(S_n, d) = n^{-1}\sum_{i,j} d(\mathrm{id}, \tau_{ij})^2$.

1) For the Hamming distance d_H it is clear that $d_H(\sigma, \sigma\tau_{ij}) = 2$, which implies $\mathrm{ObsDiam}(S_n,d_H) = 4(n-1)$. This is not sharp in many situations; for example, if we consider the function $d_H(\cdot, \mathrm{id})$, the true variance is 1 and not of order n. On the other hand, the number of fixed points is a locally Lipschitz function with respect to d_H, which converges weakly to a Poisson random variable, so that a variance estimate of order 1 in the class of Lipschitz functions cannot hold.

2) If we define for $p \in [1,\infty)$ a distance d_p on S_n by the induced ℓ^p norm

$$d_p(\sigma, \pi) = \Big(\sum_{k=1}^n |\sigma(k) - \pi(k)|^p\Big)^{\frac{1}{p}},$$

this yields $d_p(\sigma, \sigma\tau_{ij}) = 2^{1/p}|\sigma(i) - \sigma(j)|$. Consequently we have

$$\mathrm{ObsDiam}(S_n,d_p) = \frac{2^{2/p}}{6}n(n^2-1).$$

The case $p = 1$ gives Spearman's footrule and $p = 2$ Spearman's rank correlation.

3) Considering Kendall's tau, we can readily see that for two indices i,j and any $\sigma \in S_n$ it holds $I(\sigma, \sigma\tau_{ij}) \leq 2|\sigma(i) - \sigma(j)|$, since $\tau_{ij}\sigma^{-1}$ can be brought to σ^{-1} by first taking $\sigma^{-1}(i)$ to its place, and then $\sigma^{-1}(j)$. So, as above, this leads to

$$\mathrm{ObsDiam}(S_n, I) \leq \frac{2}{3}n(n^2-1).$$

4) In a more general setting, let $\rho : S_n \to \mathrm{GL}(V)$ be a faithful, unitary representation of S_n and let $\|\cdot\|$ be a unitarily invariant norm on $\mathrm{GL}(V)$. Then $d_\rho(\sigma, \tau) \coloneqq \|\rho(\sigma) - \rho(\tau)\|$ defines a bi-invariant metric on S_n, and in this case we have

$$\mathrm{ObsDiam}(S_n, d_\rho) = n^{-1}\sum_{i,j}\|\mathrm{Id} - \rho(\tau_{ij})\|.$$

5) To see the limitations of this method, consider Cayley's distance $T(\pi,\sigma)$ defined as the minimal number of transpositions required to bring π to σ. Clearly we have $T(\sigma,\sigma\tau_{ij}) = 1$ for any σ and $i \neq j$, and

$$\mathrm{ObsDiam}(S_n, T) = n - 1.$$

On the other hand, it is known that $T(\mathrm{id},\cdot)$ has variance of order $\log(n)$ (see [Dia88, Chapter 6B]), so that the variance proxy $n-1$ is grossly inaccurate.

Example 3.17. Define the random variable $f(\sigma) = S^2(\sigma,\mathrm{id}) = \sum_{i=1}^n (\sigma(i) - i)^2 = \sum_{i=1}^n (i^2 - 2i\sigma(i) + \sigma(i)^2)$. We have

$$\Gamma(f)^2(\sigma) = n^{-1} \sum_{i,j=1}^n (f(\sigma) - f(\sigma\tau_{ij}))^2 = 4n^{-1} \sum_{i,j=1}^n (\sigma(i) - \sigma(j))^2 (i-j)^2.$$

If we define the matrix $A(\sigma) = (a_{ij}(\sigma))_{i,j}$ by setting $a_{ij} = (\sigma(i) - \sigma(j))(i-j)$, then the right hand side is (up to the factor $4n^{-1}$) the squared Hilbert–Schmidt norm of $A(\sigma)$. It is clear that $|A(\sigma)|_{\mathrm{HS}} = |A(\sigma^{-1})|_{\mathrm{HS}}$, and one can also easily see that it is invariant under right multiplication with any transposition τ_{kl}. As any permutation can be written as a product of transpositions, can evaluate it at $\sigma = \mathrm{id}$. Consequently,

$$\Gamma(f)^2(\sigma) = 4n^{-1} \sum_{i,j=1}^n (i-j)^4 \leq \frac{4}{15} n^5.$$

This leads to the concentration inequality

$$\pi_n(|f - \mathbb{E}_{\pi_n} f| \geq t) \leq 2 \exp\Big(-\frac{t^2}{8n^5/15} \Big).$$

Actually, the term n^5 is natural, as the variance of f is of order n^5 (see the table above). Incorporating the variance of f into the inequality above yields

$$\pi_n(|f - \mathbb{E}_{\pi_n} f| \geq \mathrm{Var}(f)^{1/2} t) \leq 2 \exp\Big(-\frac{t^2}{19.2} \Big),$$

and so apart from the overestimated variance, it yields the correct tail behavior.

Example 3.18. Let us consider the 1-Lipschitz function $f(\sigma) = I(\sigma,\mathrm{id})$. For any $t \geq 0$ we have by (3.17) and $\mathrm{Var}_{\pi_n}(f) = n(n-1)(2n+5)/72$

$$\pi_n(|f - \mathbb{E}_{\pi_n} f| \geq \mathrm{Var}_{\pi_n}(f)^{1/2} t) \leq 2 \exp\Big(-\frac{t^2}{48} \Big),$$

which is consistent with the central limit theorem for f.

Example 3.19. We define the number of ascents $f(\sigma) = \sum_{j=1}^{n-1} \mathbb{1}_{\sigma(j+1) > \sigma(j)}$. It can be easily shown that for any $i \neq j$ the number of ascents is not sensitive to transpositions, i.e. it holds $|f(\sigma) - f(\sigma\tau_{ij})| \leq 2$. Consequently, this leads to $\Gamma(f)^2 \leq 4(n-1)$, implying the concentration inequality

$$\pi_n(|f - \mathbb{E}_{\pi_n} f| \geq t) \leq 2\exp\Big(-\frac{t^2}{8(n-1)}\Big).$$

Again, the variance term of order $\sqrt{n}$ is of the right order, as in [CKSS72] the authors have shown a central limit theorem for the number of ascents. More precisely, the sequence $g_n = (f - \mathbb{E}_{\pi_n} f)/(\sqrt{(n+1)/12})$ converges to a standard normal distribution. The above calculations lead to

$$\pi_n(|g_n| \geq t) \leq 2\exp\Big(-\frac{t^2}{96}\Big).$$

Example 3.20. A closely related statistic is given by the sum of the ascents defined as $f(\sigma) = \sum_{j=1}^{n-1}(\sigma_{i+1} - \sigma_i)_+$. A short calculation shows

$$\Gamma(f)^2 = n^{-1}\sum_{i\neq j}(f(\sigma) - f(\sigma\tau_{ij}))^2 \leq 4(n-1)^2 n^{-1}\sum_{i\neq j} = 4(n-1)^3.$$

Indeed, if we let $\Delta_{i,j} \coloneqq (\sigma(i) - \sigma(j))_+$, then

$$\begin{aligned}
&(f(\sigma) - f(\sigma\tau_{ij}))^2 \\
&= (\Delta_{i,i-1} + \Delta_{i+1,i} + \Delta_{j,j-1} + \Delta_{j+1,j} - \Delta_{j,i-1} - \Delta_{i+1,j} - \Delta_{i,j-1} - \Delta_{j+1,i})^2 \\
&\leq \max\Big(\Delta_{i,i-1} + \Delta_{i+1,i} + \Delta_{j,j-1} + \Delta_{j+1,j}, \Delta_{j,i-1} + \Delta_{i+1,j} + \Delta_{i,j-1} + \Delta_{j+1,i}\Big)^2.
\end{aligned}$$

Now each of the terms $\Delta_{i,i+1} + \Delta_{i+1,i}$, $\Delta_{j,j-1} + \Delta_{j+1,j}$ is less than $(n-1)$, and the same holds true for the two other sums. Therefore this yields

$$\pi_n(|f - \mathbb{E}_{\pi_n} f| \geq t) \leq 2\exp\Big(-\frac{t^2}{8(n-1)^3}\Big).$$

Clark [Cla09] has calculated the variance of the sum of ascents, and it is of order n^3, which is in good accordance with the concentration inequality (again, up to the factor).

Example 3.21. Finally, consider the random variable $T(\sigma) = D(\sigma) + D(\sigma^{-1})$, where $D(\sigma) = \sum_{i=1}^{n-1} \mathbb{1}_{\sigma(i+1)>\sigma(i)}$ is the number of descents. In [CD17] the authors calculated the expectation and variance of T and proven a central limit theorem. As in the above example one can easily see that $\Gamma(T)^2 \leq 4(n-1)$, as well as $\Gamma(T \circ \mathrm{inv})^2 \leq 4(n-1)$, where $\mathrm{inv} : S_n \to S_n$ denotes the inverse map. Since $\Gamma(f+g)^2 \leq 2\Gamma(f)^2 + 2\Gamma(g)^2$ holds true for any functions f,g, we also have $\Gamma(T)^2 \leq 16(n-1)$, showing that for all $t \geq 0$ it holds

$$\pi_n(|T - \mathbb{E}\, T| \geq t) \leq 2\exp\Big(-\frac{t^2}{32(n-1)}\Big).$$

Again, the variance is of order n, so that it is consistent with the CLT. Unfortunately, the constant 4 at $4n$ for $\Gamma(T)$ cannot be improved (asymptotically), which can be seen at $\widetilde{\sigma}(i) = n + 1 - i$.

3.2.4 Homogeneous polynomials in [0,1]-random variables

One possible application of Proposition 3.4 is to understand the finite n concentration properties of the so-called *d-runs on the line* as follows.

Proposition 3.22. *Let $d \in \mathbb{N}, n > d$, $(X_i)_{i=1,\ldots,n}$ be independent, identically distributed random variables with values in [0,1] and mean $\eta := \mathbb{E}\, X_1 > 0$. Define the random variable $f_d(X) := \sum_{i=1}^n X_i \cdots X_{i+d-1}$, where the indices are to be understood modulo n. For any $t \geq 0$ it holds*

$$\mathbb{P}\Big(f_d(X) - \mathbb{E}\, f_d(X) \geq n^{1/2}\eta^{d/2} t \Big) \leq 2 \exp\Big(- \frac{t^2}{2d^2(1 + n^{-1/2}\eta^{-d/2} t)} \Big). \tag{3.18}$$

If we consider d to be fixed, the bounded difference inequality provides wrong bounds for $f_d(X)$, as we have $|f_d(X) - f_d(\overline{X}_i, X_i')| \leq d$ for any $i \in [n]$, yielding

$$\mathbb{P}(|f_d(X) - \mathbb{E}\, f_d(X)| \geq t) \leq 2 \exp\Big(- \frac{2t^2}{nd^2} \Big).$$

This suggests of a normalization of $f(X)$ by $n^{-1/2}$. On the other hand, in [RR09, Theorem 4.1], the authors prove a CLT for the d-runs on the line for Bernoulli random variables X_i with success probability p, by normalizing f by $p^{-d/2}n^{-1/2}$. This is also the reason for the choice $n^{1/2}\eta^{d/2}t$ in inequality (3.18). In other words, under the assumption $n\eta^d \to \infty$ as $n \to \infty$, this yields sub-Gaussian tails for $n^{-1/2}\eta^{-d/2} f$. This is in good accordance with the CLT proven in [RR09, Theorem 4.1].

Example 3.23. If $(X_v)_{v \in E(K_n)}$ is the Erdös–Rényi model with parameter p, for any fixed graph H with v vertices and e edges, the subgraph counting statistic T_H can be written in the form (3.7) with $w_e = 1$, and $k = e$. Furthermore, it is easy to see that $\mathrm{ML}(f) \leq n^{\Delta - 1}$ for the maximum degree Δ, so that Corollary 0.2 yields

$$\mathbb{P}(T_H(X) - \mathbb{E}\, T_H(X) \geq \varepsilon\, \mathbb{E}\, T_H(X)) \leq \exp\Big(- C_{k,\varepsilon} n^{v - \Delta + 1} p^e \Big).$$

For example, this gives nontrivial bounds in the triangle case whenever $n^2p^3 \to \infty$ as $n \to \infty$. This bound is clearly suboptimal, and the optimal decay is known, see [Cha12; DK12]. However, this is better than the bound obtained by the bounded differences inequality. In general, if we consider subgraph counting statistics for some subgraph H with v vertices and e edges on an Erdös–Rényi model $(X_v)_{v \in E(K_n)}$, the bounded difference inequality yields the estimate

$$\mathbb{P}(f(X) - \mathbb{E}\, f(X) \geq \varepsilon\, \mathbb{E}\, f(X)) \leq \exp\Big(- C_{\varepsilon,H} \frac{n^{2v}p^{2e}}{n^2 n^{2\Delta - 2}} \Big).$$

Thus, to obtain non-trivial estimates in the limit $n \to \infty$, one has to assume that $n^{v-\Delta}p^e \to \infty$. With the above inequality, this can be weakened to $nn^{v-\Delta}p^e \to \infty$.

Example 3.24. Let $(X_{ij})_{i,j=1,\dots,n}$ be a random matrix with independent and [0,1]-valued entries X_{ij}, and let $a = (a_i)_{i=1,\dots,n}$ be a vector with positive entries. Consider the random variable

$$f(X) \coloneqq \max_{\pi \in S_n} \sum_{i=1}^n a_i X_{i,\pi(i)}.$$

(For a complete bipartite graph $G = (V,E)$ with vertex set $V = V_0 \cup V_1$ for $V_0 = \{1,\dots,n\}$ and $V_1 = \{n+1,\dots,n\}$, equipped with the weights $w_{ij} = a_i X_{i,j-n}$, $f(X)$ is the value of the matching of G maximizing the total weight.)

Now, for any pair (i,j) set $f_{ij}(x_{\{ij\}^c}) \coloneqq f(x_{(i,j)^c}, 0)$. For a fixed $x \in [0,1]^V$ let $\widetilde{\pi}$ be the permutation attaining the maximum and observe that

$$\begin{aligned}\sum_{i,j=1}^n (f(X) - f_{ij}(X_{(i,j)^c}))^2 &\leq \sum_{i,j=1}^n \Big(\sum_{k=1}^n a_k (X_{k,\widetilde{\pi}(k)} - (X_{(i,j)^c},0)_{k,\widetilde{\pi}(k)}) \Big)^2 \\ &\leq \sum_{i,j=1}^n \mathbb{1}_{j=\widetilde{\pi}(i)} a_i^2 X_{i,\widetilde{\pi}(i)}^2 \leq \|a\|_\infty f(X).\end{aligned}$$

Consequently, [BLM09, Theorem 1] yields

$$\mathbb{P}(f(X) - \mathbb{E}\, f(X) \geq t) \leq \exp\Big(- \frac{t^2}{2\|a\|_\infty (\mathbb{E}\, f(X) + t/2)} \Big).$$

3.3 Proofs and auxiliary results

We begin by proving Theorem 3.1. Before we start, let us recall a result proven in [BG99, Theorem 2.1], relating the exponential moments of $f - \mathbb{E}_\mu f$ to those of $\Gamma(f)^2$.

Theorem 3.25. *Assume that (Ω, μ, Γ) satisfies the $\Gamma-$mLSI(ρ) (3.2) with constant $\rho > 0$. Then for any $f \in \mathcal{A}$ and any $\alpha > \frac{\rho}{2}$ we have*

$$\mathbb{E}_\mu \exp(f - \mathbb{E}_\mu f) \leq \big(\mathbb{E}_\mu \exp(\alpha \Gamma(f)^2) \big)^{\frac{\rho}{2\alpha - \rho}}.$$

Furthermore, we need an elementary inequality to adjust the constants in concentration or deviation inequalities: for any two constants $c_1 > c_2 > 1$ we have for all $r \geq 0$ and $c > 0$

$$c_1 \exp(-cr) \leq c_2 \exp\Big(- \frac{\log(c_2)}{\log(c_1)} cr \Big) \tag{3.19}$$

whenever the left hand side is smaller or equal to 1.

Lastly, for the proof of Proposition 3.2 we require a lemma which provides concentration inequalities for random variables with a certain growth condition of their moment generating function. Define $h(x) \coloneqq 1 + x - \sqrt{1+2x} = \frac{1}{2}(1 - \sqrt{1+2x})^2$.

Lemma 3.26 (Lemma 11 in [BLM03]). *Let C and a denote two positive real numbers. Then*

$$\sup_{\lambda \in [0,1/a)} \Big(\lambda t - \frac{C\lambda^2}{1 - a\lambda} \Big) = \frac{2C}{a^2} h\Big(\frac{at}{2C} \Big) \geq \frac{t^2}{2(2C + at/3)}.$$

Proof of Theorem 3.1. First, by making use of Theorem 3.25 in the first and $a^2 \leq 2(a-b)_+^2 + 2b^2$ for any $a,b \geq 0$ in the third step, for all $\lambda \geq 0$ we obtain

$$\begin{aligned} \mathbb{E}_\mu \exp\big(\lambda(f - \mathbb{E}_\mu f)\big) &\leq \mathbb{E}_\mu \exp\big(\lambda^2 \rho \Gamma(f)^2\big) \leq \mathbb{E}_\mu \exp\big(\lambda^2 \rho g^2\big) \\ &\leq \exp\big(2\lambda^2 \rho (\mathbb{E}_\mu g)^2\big) \, \mathbb{E}_\mu \exp\big(2\lambda^2 \rho (g - \mathbb{E}_\mu g)_+^2\big). \end{aligned}$$

To estimate the right hand side further, we apply Theorem 3.25 to $f := \lambda g$, use Markov's inequality and optimize to show that for all $t \geq 0$

$$\mu(g - \mathbb{E}_\mu g \geq t) \leq \exp\Big(- \frac{t^2}{2\rho b^2} \Big). \tag{3.20}$$

Here, to obtain the factor 2 in the denominator, one has to let $\alpha \to \infty$ in Theorem 3.25. Thus, if we define $h \coloneqq 2\lambda^2 \rho (g - \mathbb{E}_\mu g)_+^2$ this leads to

$$\mathbb{E}_\mu \exp(h) \leq 1 + \int_0^\infty \exp\Big(- t\Big(\frac{1}{4\rho^2 b^2 \lambda^2} - 1 \Big) \Big) dt = \frac{1}{1 - 4\rho^2 b^2 \lambda^2}$$

if $\lambda^2 < 1/(4\rho^2 b^2)$. Let us set $c \coloneqq 2\rho b$ and $a^2 \coloneqq (\mathbb{E}_\mu g)^2$. Consequently, for all $\lambda \in [0, c^{-1})$ we have by Markov's inequality

$$\mu(f - \mathbb{E}_\mu f \geq t) \leq \exp\Big(- \lambda t + 2\lambda^2 \rho a^2 \Big) \frac{1}{1 - \lambda^2 c^2}. \tag{3.21}$$

Now we distinguish the two cases $(i) : t \leq a^2 b^{-1}$ and $(ii) : t > a^2 b^{-1}$. In the first case, set $\lambda \coloneqq \frac{t}{4\rho a^2}$ (which implies $\lambda^2 c^2 \leq 1/4$ and thus is in the range) to obtain

$$\exp\big(- \lambda t + 2\lambda^2 \rho a^2 \big) \frac{1}{1 - \lambda^2 c^2} \leq \frac{4}{3} \exp\Big(- \frac{t^2}{4\rho a^2} + \frac{t^2}{8\rho a^2} \Big) = \frac{4}{3} \exp\Big(- \frac{t^2}{8\rho a^2} \Big), \tag{3.22}$$

where we have used that $\frac{1}{1-x}$ is increasing and is less than $4/3$ for $x \leq 1/4$. In the second case, we set $\lambda \coloneqq \frac{1}{4\rho b}$ (which is equivalent to $\lambda^2 c^2 = 1/4$) to arrive at

$$\exp\big(- \lambda t + 2\lambda^2 \rho a^2 \big) \frac{1}{1 - \lambda^2 c^2} \leq \frac{4}{3} \exp\big(- \frac{t}{4\rho b} + \frac{t}{8\rho b} \big) \leq \frac{4}{3} \exp\big(- \frac{t}{8\rho b} \big). \tag{3.23}$$

Combining (3.22) and (3.23) finishes the proof of (3.3). From here, (3.4) follows by considering $-f$ instead of f and using (3.19) to change the constant 8/3 to 2.

Finally, the Bernstein-type inequalities can be easily derived from (3.3) and (3.4) respectively. Indeed, a short calculation shows that for all $a, b > 0$ and $t \geq 0$ we have

$$\frac{t^2}{a^2+bt} \leq \min\Big(\frac{t^2}{a^2}, \frac{t}{b}\Big) = \frac{t^2}{\max(a^2,bt)} \leq \frac{2t^2}{a^2+bt}.$$

□

Proof of Proposition 3.2. Again, the proof is based on Theorem 3.25. Choosing $\alpha = \rho$, applying the exponential inequality to λf and using the monotonicity leads to

$$\mathbb{E}_\mu \exp\big(\lambda(f - \mathbb{E}_\mu f)\big) \leq \exp\big(\lambda^2\rho(b + a\,\mathbb{E}_\mu f)\big)\,\mathbb{E}_\mu \exp\big(\lambda^2\rho a(f - \mathbb{E}_\mu f)\big).$$

Thus for $\lambda \in (0,(a\rho)^{-1})$, by Jensen's inequality applied to the concave function $x \mapsto x^{\lambda\rho a}$ we have

$$\big(1 - \lambda\rho a\big)\log\big(\mathbb{E}_\mu \exp\big(\lambda(f - \mathbb{E}_\mu f)\big)\big) \leq \lambda^2\rho(b + a\,\mathbb{E}_\mu f).$$

Finally, Markov's inequality and Lemma 3.26 yield the claim.

For the lower bound, assume without loss of generality $\rho = 1$, which can always be achieved by defining $\Gamma_\rho := \sqrt{\rho}\Gamma$. f is (a,b)-self-bounding with respect to Γ if and only if it is $(\rho a, \rho b)$-self-bounding with respect to Γ_ρ. Now by the self-boundedness assumption, we have for any $\lambda > 0$

$$\begin{aligned}\mathbb{E}\exp\big(\lambda(\mathbb{E} f - f)\big) &\leq \mathbb{E}\exp\big(\Gamma(-\lambda f)^2\big) \leq \mathbb{E}\exp\big(\lambda^2\Gamma(f)^2\big) \leq \mathbb{E}\exp\big(\lambda^2(af + b)\big)\\ &\leq \exp\big(\lambda^2(a\,\mathbb{E} f + b)\big)\,\mathbb{E}\exp\big(\lambda^2 a(f - \mathbb{E} f)\big).\end{aligned}$$

We estimate the second factor as follows: For $\lambda a < 1$, concavity of the function $x \mapsto x^{\lambda a}$ and the estimate from the first part give

$$\mathbb{E}\exp\big(\lambda^2 a(f - \mathbb{E} f)\big) \leq \big(\mathbb{E}\exp\big(\lambda(f - \mathbb{E} f)\big)\big)^{\lambda a} \leq \exp\Big(\lambda a\frac{\lambda^2}{1-\lambda a}(a\,\mathbb{E} f + b)\Big).$$

Combining these estimates yields

$$\mathbb{E}\exp\big(\lambda(\mathbb{E} f - f)\big) \leq \exp\Big(\frac{\lambda^2}{1-\lambda a}(a\,\mathbb{E} f + b)\Big),$$

from which the concentration inequalities follow from Lemma 3.26 as above. □

Proof of Proposition 3.3. The proof is a slight modification of the proof for independent random variables in [BLM03]. As stated in Proposition 3.14, the uniform measure π_n on S_n satisfies a Γ_+-mLSI(2) with respect to

$$\Gamma_+(f)^2(\sigma) = \frac{1}{n}\sum_{i,j=1}^n (f(\sigma) - f(\sigma\tau_{ij}))_+^2.$$

Writing $f_A(\sigma) \coloneqq d_T(\sigma, A)$, it is well known (see [BLM03]) that we have

$$f_A(\sigma) = \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha \in \mathbb{R}^n : |\alpha|_2 = 1} \sum_{k=1}^n \alpha_k \nu(\sigma' : \sigma'_k \neq \sigma_k),$$

where $\mathcal{M}(A)$ is the set of all probability measures on A. To estimate $\mathit{\Gamma}_+(f_A)^2(\sigma)$, one has to compare $f_A(\sigma)$ and $f_A(\sigma\tau_{ij})$. To this end, for any $\sigma \in S_n$ fixed, let $\widetilde{\alpha}, \widetilde{\nu}$ be parameters maximizing $f_A(\sigma)$, and let $\hat{\nu} = \hat{\nu}_{ij}$ be a minimizer of $\inf_{\nu \in \mathcal{M}(A)} \sum_{k=1}^n \widetilde{\alpha}_k \nu(\sigma' : \sigma'_k \neq (\sigma\tau_{ij})_k)$. This leads to

$$\begin{aligned}
\mathit{\Gamma}_+ f_A(\sigma)^2 &\leq \frac{1}{n} \sum_{i,j=1}^n \big(\sum_{k=1}^n \widetilde{\alpha}_k (\hat{\nu}(\sigma'_k \neq \sigma_k) - \hat{\nu}(\sigma'_k \neq (\sigma\tau_{ij})_k))\big)_+^2 \\
&= \frac{1}{n} \sum_{i,j=1}^n \big(\widetilde{\alpha}_i(\hat{\nu}(\sigma'_i \neq \sigma_i) - \hat{\nu}(\sigma'_i \neq \sigma_j)) + \widetilde{\alpha}_j(\hat{\nu}(\sigma'_j \neq \sigma_j) - \hat{\nu}(\sigma'_j \neq \sigma_i))\big)_+^2 \\
&\leq \frac{2}{n} \sum_{i,j=1}^n (\widetilde{\alpha}_i^2 + \widetilde{\alpha}_j^2) \leq 4.
\end{aligned}$$

Hence, by similar arguments as in the proof of Theorem 3.1 we have for any $\lambda \geq 0$

$$\mathbb{E}_{\pi_n} \exp\big(\lambda(f_A - \mathbb{E}_{\pi_n} f_A)\big) \leq \exp(4\lambda^2), \tag{3.24}$$

implying the sub-Gaussian estimate

$$\pi_n(f_A - \mathbb{E}_{\pi_n} f_A \geq t) \leq \exp\Big(-\frac{t^2}{16}\Big).$$

Fix a set $A \subseteq S_n$ satisfying $\pi_n(A) \geq 1/2$. As a $\mathit{\Gamma}_+-$mLSI(2) implies a Poincaré inequality (see [BT06, Proposition 3.5], or Lemma 2.2 for a similar reasoning), we also have (by Chebyshev's inequality)

$$t^2 \pi_n\big(f_A - \mathbb{E}_{\pi_n} f_A \leq -t\big) \leq \mathrm{Var}_{\pi_n}(f_A) \leq \mathbb{E}_{\pi_n} \mathit{\Gamma}_+(f_A)^2 \leq 4,$$

which evaluated at $t = \mathbb{E}_{\pi_n} f_A$ yields $(\mathbb{E}_{\pi_n} f_A)^2 \leq 8$. Thus, for any $t \geq \sqrt{8}$ it holds

$$\pi_n(f_A \geq t) \leq \exp\Big(-\frac{(t-\sqrt{8})^2}{16}\Big) \leq \sqrt{e}\exp\Big(-\frac{t^2}{32}\Big) \leq 2\exp\Big(-\frac{t^2}{32}\Big), \tag{3.25}$$

where the second-to-last inequality follows from $(t - \sqrt{8})^2 \geq t^2/2 - 8$ for any $t \geq \sqrt{8}$. For $t \leq \sqrt{8}$ the inequality (3.25) holds trivially. $\square$

Proof of Proposition 3.4. We show that f is weakly (kML(f),0)-self bounding in the language of [BLM09]. To see this, for any $v \in V$ let $f_v(x_{v^c}) \coloneqq \sum_{e \in E : v \notin E} w_e X_e =$

$f(X_{v^c}, 0)$. Now we have

$$\sum_{v\in V}(f(x) - f_v(x_{v^c}))^2 = \sum_{v\in V}\Big(X_v \sum_{e\in E: v\in E} w_e X_{e\setminus v}\Big)^2 \le \sum_{v\in V} X_v \partial_v f(X)^2$$
$$\le \mathrm{ML}(f) \sum_{v\in V} X_v \partial_v f(X) = k\mathrm{ML}(f) f(X).$$

Here, the first inequality follows from $X_v \in [0,1]$ and the last step is a consequence of Euler's homogeneous function theorem. Consequently, [BLM09, Theorem 1] yields for any $t \ge 0$

$$\mathbb{P}(f(X) - \mathbb{E}\, f(X) \ge t) \le \exp\Big(-\frac{t^2}{2k\mathrm{ML}(f)(\mathbb{E}\, f(X) + t/2)}\Big).$$

For the lower bound, apply [BLM09, Theorem 1] on $\widetilde{f} = \mathrm{ML}(f)^{-1} f$ which satisfies $0 \le \widetilde{f}(x) - \widetilde{f}_v(x_{v^c}) \le 1$ for all $v \in V$ and $x \in [0,1]^V$ and is weakly $(k\mathrm{ML}(f)^{-1},0)$-self bounding. □

Proof of Proposition 3.6. The first part follows as above. As for the second part, if we choose $\mathcal{F} = \mathcal{F}_q = \{a \in \mathbb{R}^V : a_v \ge 0, \|a\|_q \le 1\}$ for some $q \in [1,\infty]$ this leads to

$$f_{\mathcal{F}}(X) = \sup_{a\in\mathcal{F}_q} \sum_{v\in V} a_v X_v = \Big(\sum_{v\in V} |X_v|^p\Big)^{1/p}$$

for the Hölder conjugate p, which is due to the nonnegativity of the X_i and the dual formulation of the L^p norm in $\mathbb{R}^V$. □

Remark. Actually, in the case $a_i \in [0,1]^n$ we can also invoke [BLM00, Theorem 2.1] to obtain Poisson tail bounds for f. Indeed, if we set $Z = f(X)$ and $Z^{(v)} = f(X_{v^c})$, it is easy to verify $0 \le Z - Z^{(v)} \le 1$ as well as $\sum_{v\in V}(Z - Z^{(v)}) \le Z$, so that [BLM00] can be applied, yielding

$$\mathbb{P}(Z - \mathbb{E}\, Z \ge t) \le \exp\Big(t - (\mathbb{E}\, Z + t)\log\Big(1 + \frac{t}{\mathbb{E}\, Z}\Big)\Big).$$

For similar calculations, see [BLM00, remark (1) on page 281].

Proof of Proposition 3.7. Inequality (3.11) follows immediately from Theorem 3.1 and the inequality

$$|\nabla_S|\nabla_S f|| \le \|f_S''\|_{\mathrm{op}},$$

proven in [BCG17, Lemma 3.1]. Now, if f is orthogonal to all affine functions (in $L^2(\sigma_{n-1})$), [BCG17, Proposition 5.1] shows

$$\mathbb{E}_{\sigma_{n-1}} |\nabla_S f|^2 \le \rho_n\, \mathbb{E}_{\sigma_{n-1}} \|f_S''\|_{\mathrm{HS}}^2,$$

from which (3.12) can readily deduced. □

Proof of Proposition 3.8. The proof is similar to the proof of Proposition 3.7. It only remains to see that $|\nabla|\nabla f|| \leq \|f''\|_{\mathrm{op}}$, cf. [GS20, Lemma 7.2], and argue as above. $\square$

Proof of Corollary 3.9. This is an immediate consequence of Theorem 3.1 and the Poincaré inequality upon noticing that for $g(y) := f(y) - \mathbb{E}_\mu f(x) - \langle y - \mathbb{E}_\mu(x), \mathbb{E}_\mu \nabla f(x)\rangle$ we have $\nabla g = \nabla f - \mathbb{E}_\mu \nabla f$ and $g'' = f''$. $\square$

Proof of Proposition 3.10. The idea of the proof of the mLSIs is already present in [BG07]. Let $(\Omega, \mathcal{F}, \nu)$ be any probability space. For any function g we have due to the inequality $(a-b)_+(e^a - e^b)_+ \leq \frac{1}{2}(a-b)_+^2(e^a + e^b) \leq (a-b)_+^2 e^a$

$$\mathrm{Cov}_\nu(g, e^g) \leq \iint (g(x) - g(y))_+^2 d\nu(y) e^{g(x)} d\nu(x).$$

Applying this to $\nu = \mu(\cdot \mid \overline{x}_i)$ and $g = f(\overline{x}_i, \cdot)$ and using (3.13) yields

$$\mathrm{Ent}_\mu(e^f) \leq \sigma^2 \sum_{i=1}^n \iint (f(x) - f(\overline{x}_i, x_i'))_+^2 d\mu(x_i' \mid \overline{x}_i) e^{f(x)} d\mu(x) = \sigma^2 \int |\mathfrak{d}^+ f|^2 e^f d\mu.$$

The second inequality follows by similar reasoning, by observing that

$$\iint (g(x) - g(y))_+^2 (e^{g(x)} + e^{g(y)}) d\nu(x) d\nu(y) = \int e^{g(x)} \int (g(x) - g(y))^2 d\nu(y) d\nu(x).$$

The exponential inequalities are a consequence of Theorem 3.25. $\square$

For the next proofs, recall the duality formula $|x| = \sup_{y \in S^{n-1}} \langle x, y\rangle$.

Proof of Proposition 3.11. Let us bound $|\mathfrak{d}^+ h|^2$. Fix x, choose some $\widetilde{A}$ maximizing $\sup_{A \in \mathcal{A}} \langle x, Ax\rangle$ and use the monotonicity of $y \mapsto y_+$ to obtain

$$\begin{aligned}
|\mathfrak{d}^+ h(x)| &= \Big(\sum_{i=1}^n \int (h(x) - h(\overline{x}_i, x_i'))_+^2 d\mu(x_i' \mid \overline{x}_i)\Big)^{1/2} \\
&\leq \Big(\sum_{i=1}^n \sup_{x_i'} \Big(2(x_i - x_i') \sum_{j=1}^n \widetilde{A}_{ij} x_j\Big)_+^2\Big)^{1/2} \\
&\leq 4\|\widetilde{A}x\|_2 \leq 4 f_{\mathcal{A}}(x).
\end{aligned}$$

Now, for some maximizer $\widetilde{A}$ of $\sup_{A \in \mathcal{A}} \|Ax\|$ and $\widetilde{v}$ for $\sup_{v \in S^{n-1}} \langle \widetilde{A}x, v\rangle$ it holds

$$\begin{aligned}
|\mathfrak{d}^+ f_{\mathcal{A}}|^2 &\leq \sum_i \sup_{x_i'} \big(\sup_v \langle \widetilde{A}x, v\rangle - \sup_v \langle \widetilde{A}(\overline{x}_i, x_i'), v\rangle\big)_+^2 \\
&\leq \sum_i \sup_{x_i'} \big((x_i - x_i')\langle \widetilde{A}e_i, \widetilde{v}\rangle\big)_+^2 \leq 4 \sum_i \langle \widetilde{A}e_i, \widetilde{v}\rangle^2 \\
&\leq 4\big(\sup_w \sum_i w_i \langle \widetilde{A}e_i, \widetilde{v}\rangle\big)^2 \leq 4 \sup_{A \in \mathcal{A}} \|A\|_{\mathrm{op}}^2.
\end{aligned}$$

Here, the suprema of v and w are taken over the n-dimensional sphere. We can now apply Theorem 3.1 with $\varGamma = \mathfrak{d}^+, \rho = 2\sigma^2$, $g = 4f_{\mathcal{A}}$ and $b = 8\varSigma$ to finish the proof. □

Proof of Corollary 3.12. The inequality $(f(x) - f(y))_+^2 \leq f'(x)^2(x-y)^2$ holds for any convex function f, which readily implies

$$|\mathfrak{d}^+ f|^2 \leq \sum_{i=1}^n \partial_i f(x)^2 \int (x_i - x_i')^2 d\mu(x_i' \mid \overline{x}_i) \leq 4|\nabla f|^2.$$

The statements now follow from Proposition 3.10. □

Proof of Proposition 3.13. In the notation of Theorem 3.1, by Corollary 3.12 and its proof we can take $g \coloneqq 2|\nabla f|$ (if $\varGamma = \mathfrak{d}^+$). For fixed $x \in [-1, +1]^n$ and a $y \in S^{n-1}$ such that $|\nabla f|(x) = \langle y, \nabla f(x)\rangle$ we have by the mean value theorem

$$|\mathfrak{d}^+ |\nabla f|| \leq \sum_i \sup_{x_i'} \big(\langle y, \nabla f(x) - \nabla f(\overline{x}_i, x_i')\rangle\big)_+ \leq 2 \sup_{y,z \in S^{n-1}} \sum_{i,j} y_i z_j c_{ij} = 2\|c\|_{\mathrm{op}}.$$

Now apply Theorem 3.1 to $\varGamma = \mathfrak{d}^+$, $\rho = 2\sigma^2$, $g \coloneqq 2|\nabla f|$ and $b = 4\|c\|_{\mathrm{op}}$. □

Proof of Proposition 3.14. Using and rewriting [GQ03, Theorem 1] we obtain for any $f : S_n \to \mathbb{R}$

$$\mathrm{Ent}(e^f) \leq \frac{1}{2nn!} \sum_{i,j=1}^n \sum_{\sigma \in S_n} (f(\sigma\tau_{ij}) - f(\sigma))(e^{f(\sigma\tau_{ij})} - e^{f(\sigma)}).$$

Now, using the inequality $(a-b)(e^a - e^b) \leq \frac{1}{2}(e^a + e^b)(a-b)^2$ and the fact that $\sigma \mapsto \sigma\tau_{ij}$ is an automorphism of S_n leads to $\varGamma-$mLSI(1). The $\varGamma_+-$mLSI(2) follows in the same manner from the inequality $(a-b)_+(e^a - e^b) \leq (a-b)_+^2 e^a$. □

Proof of Proposition 3.22. Clearly, f_d is d-homogeneous and has positive weights in the sense of (3.7), if we set $V = [n]$ and $E = \{\{j, j+1, \ldots, j+d-1\}, j = 1, \ldots, n\}$, $w_e = 1$. Furthermore, the partial derivatives can be easily bounded: For any fixed $l \in [n]$ there are exactly d terms which depend on X_l, and the product is bounded by 1. Consequently, $\mathrm{ML}(f_d) = \max_{l \in [n]} \max_{x \in [0,1]^n} \partial_l f(X) = d$. Thus, Proposition 3.4 yields for all $t \geq 0$

$$\mathbb{P}(f_d(X) - \mathbb{E}\, f_d(X) \geq t) \leq \exp\Big(-\frac{t^2}{2d^2(\mathbb{E}\, f_d(X) + t/2)}\Big).$$

The assertion follows easily from $\mathbb{E}\, f_d(X) = n\eta^d$. □

CHAPTER 4

Concentration inequalities for bounded functions of independent and weakly dependent random variables

The aim of this chapter is to prove concentration of measure results for general bounded functions in independent and weakly dependent random variables. In the first section we formulate concentration inequalities in these two settings. Afterwards, in Section 4.2, we provide various applications, such as

1. deviation inequalities for functions of suprema type of the form $f(X) = \sup_{t\in\mathcal{T}}\|\sum_{i_1\neq\ldots\neq i_d} t_{i_1\ldots i_d} X_{i_1}\cdots X_{i_d}\|$ for a compact set of vectors $\mathcal{T}$ from some Banach space $(\mathcal{B}, \|\cdot\|)$,

2. concentration properties of U-statistics $f(X) = \sum_{i_1\neq\ldots\neq i_d} h(X_{i_1},\ldots,X_{i_d})$ for bounded kernels h,

3. concentration inequalities for polynomials in the Ising model.

All the results will be proven in Section 4.3.

In this and the next chapter we choose to work with random variables instead of measures. By abuse of language we say that a random vector $X = (X_1,\ldots,X_n)$ satisfies a $\Gamma-\mathrm{LSI}(\sigma^2)$ for some difference operator Γ, if its distribution does.

4.1 General results

The first theorem provides general concentration inequalities for bounded functions of independent random variables $X_1,\ldots,X_n$ or a random vector $X = (X_1,\ldots,X_n)$ satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. Recall that $\mathcal{Y}$ is some generic product space in the independent case and a product of finite spaces for dependent random variables. Furthermore, recall the definition of the higher order difference operators $\mathfrak{h}^{(k)}$ from Section 2.2 and the definitions of the operator norms from Section 2.1.

Theorem 4.1. *Let $d\in\mathbb{N}$, $X = (X_1,\ldots,X_n)$ be a random vector and $f:\mathcal{Y}\to\mathbb{R}$ a measurable function satisfying $f = f(X)\in L^\infty(\mathbb{P})$.*

1. *If $X_1, \ldots, X_n$ are independent random variables, we have for any $t \geq 0$*

$$\begin{aligned} &\mathbb{P}\left(|f - \mathbb{E} f| \geq t\right) \\ &\leq 2 \exp\Big(- \frac{1}{217} \min_{k \in [d-1]} \Big(\frac{t}{d \|\mathfrak{h}^{(k)} f\|_{\mathrm{op},1}} \Big)^{\frac{2}{k}} \wedge \Big(\frac{t}{d \|\mathfrak{h}^{(d)} f\|_{\mathrm{op},\infty}} \Big)^{\frac{2}{d}} \Big). \end{aligned} \tag{4.1}$$

2. *If X satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$, then for any $t \geq 0$ it holds*

$$\begin{aligned} &\mathbb{P}\left(|f - \mathbb{E} f| \geq t\right) \\ &\leq 2 \exp\Big(- \frac{1}{11\sigma^2} \min_{k \in [d-1]} \Big(\frac{t}{d \|\mathfrak{h}^{(k)} f\|_{\mathrm{op},1}} \Big)^{\frac{2}{k}} \wedge \Big(\frac{t}{d \|\mathfrak{h}^{(d)} f\|_{\mathrm{op},\infty}} \Big)^{\frac{2}{d}} \Big). \end{aligned} \tag{4.2}$$

Here, the order d can be chosen freely and depending on the situation at hand, as the differences of order d need to be bounded uniformly. For example, we use Theorem 4.1 to prove concentration inequalities for U-statistics in Proposition 4.8 and for polynomials in the Ising model in Theorem 4.9. In both cases, d will be equal to the order of the function.

The special case $d = 2$ can be considered as a generalized form of the Hanson–Wright inequality for any bounded function f. Indeed, for $d = 2$ we can write (4.1) and (4.2) as

$$\mathbb{P}(|f - \mathbb{E} f| \geq t) \leq 2 \exp\Big(- \frac{1}{C} \min\Big(\frac{t^2}{\|\mathfrak{h} f\|_{\mathrm{op},1}^2}, \frac{t}{\|\mathfrak{h}^{(2)} f\|_{\mathrm{op},\infty}} \Big)\Big).$$

In Chapter 5 we prove a Hanson–Wright-type inequality for α-sub-exponential random variables which allows to treat unbounded random variables as well. Theorem 4.1 produces Hanson–Wright-type inequalities for bounded X_i only.

Using straightforward estimates for the norms of d-tensors (i. e. $|A|_{\mathrm{op}} \leq |A|_{\mathrm{HS}}$ for any d-tensor A), the following corollary easily follows, which in particular allows to recover the results in [GSS19b]. As this is an immediate consequence of the above inequality and Jensen's inequality, we omit the proof.

Corollary 4.2. *Let $d \in \mathbb{N}$, $X = (X_1, \ldots, X_n)$ be a random vector satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ and $f = f(X) \in L^\infty(\mathbb{P})$. For any $t \geq 0$ we have*

$$\mathbb{P}(|f - \mathbb{E} f| \geq t) \leq 2 \exp\Big(- \frac{1}{11\sigma^2} \min_{k \in [d-1]} \Big(\frac{t}{d \|\mathfrak{h}^{(k)} f\|_{\mathrm{HS},2}} \Big)^{\frac{2}{k}} \wedge \Big(\frac{t}{d \|\mathfrak{h}^{(d)} f\|_{\mathrm{HS},\infty}} \Big)^{\frac{2}{d}} \Big).$$

Next, we prove concentration inequalities for functions in weakly dependent random variables which resemble multilinear polynomials, constructed in the following way. Suppose that $\mathcal{Y} = \mathcal{X}^{\mathcal{I}}$ for some finite sets $\mathcal{X}$ and $\mathcal{I}$ equipped with some probability measure μ. Let $f : \mathcal{X} \to \mathbb{R}$ and A be a d-tensor with vanishing diagonal. For any $J \subset \mathcal{I}$ we write

$$f_J(y) = \prod_{i \in J} f(y_i) \quad \text{and} \quad \widetilde{f}_J(y) = \prod_{i \in J} \Big(f(y_i) - \int f(y_i) d\mu \Big)$$

and introduce the short-hand notations

$$\mu_J := \mathbb{E}_\mu f_J \quad \text{and} \quad \widetilde{\mu}_J := \mathbb{E}_\mu \widetilde{f}_J.$$

Recall that for any finite set $\mathcal{J}$ we denote by $\mathcal{P}(\mathcal{J})$ the set of all partitions of $\mathcal{J}$. We let $N : \mathcal{P}(\mathcal{J}) \to \mathbb{N}_0$ be the number of singletons in a partition P (i. e. the number of sets $\{i_j\}, i_j \in \mathcal{J}$), and $M : \mathcal{P}(\mathcal{J}) \to \mathbb{N}_0$ the number of subsets with more than one element. Finally, we define a polynomial-like function $f_{d,A} : \mathcal{Y} \to \mathbb{R}$ as

$$f_{d,A} = \sum_{I \in \mathcal{I}^d} A_I \sum_{P \in \mathcal{P}(I)} g_P := \sum_{I \in \mathcal{I}^d} A_I \sum_{P \in \mathcal{P}(I)} (-1)^{M(P)} \prod_{\substack{J \in P \\ |J|=1}} \widetilde{f}_J \prod_{\substack{J \in P \\ |J|>1}} \widetilde{\mu}_J. \tag{4.3}$$

Note that if $\mathcal{X} \subset \mathbb{R}$, one choice of the spin function is the identity $f(x) = x$, for which $f_{d,A}$ becomes a multilinear polynomial in $x_1, \ldots, x_n$.

Proposition 4.3. *Let $d \in \mathbb{N}$, $X = (X_1, \ldots, X_n)$ be a random vector satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$, A a d-tensor with vanishing diagonal and $f : \mathcal{X} \to \mathbb{R}$ be such that $\sup_{x,y}|f(x) - f(y)| \le c$. Then, $f_{d,A} = f_{d,A}(X)$ as in (4.3) is a centered random variable, for all $p \ge 2$ we have*

$$\|f_{d,A}\|_p \le \sigma^d c^d |A|_{\mathrm{HS}}\, p^{d/2}, \tag{4.4}$$

and consequently for any $t \ge 0$ it holds

$$\mathbb{P}(|f_{d,A}| \ge t) \le 2 \exp\Big(- \frac{1}{22\sigma^2 c^2} \Big(\frac{t}{|A|_{\mathrm{HS}}} \Big)^{\frac{2}{d}} \Big). \tag{4.5}$$

Proposition 4.3 will be used to prove statements about the number of triangles in exponential random graph models, see Theorem 4.11, Corollary 4.12 and Proposition 4.13. See also the discussion of the polynomials in the Ising model with an external field after Corollary 4.10.

4.2 Applications

4.2.1 Deviation inequalities for empirical processes

First we consider a uniform version of polynomial chaos. Let $\mathcal{I}_{n,d}$ denote the family of subsets of $[n]$ with d elements, fix a Banach space $(\mathcal{B}, \|\cdot\|)$ with its dual space $(\mathcal{B}^*, \|\cdot\|_*)$, a compact subset $\mathcal{T} \subset \mathcal{B}^{\mathcal{I}_{n,d}}$ and let $\mathcal{B}_1^*$ be the 1-ball in $\mathcal{B}^*$ with respect to $\|\cdot\|_*$. Let $X = (X_1, \ldots, X_n)$ be a random vector with support in $[-1, +1]^n$ and define

$$f(X) := f_{\mathcal{T}}(X) := \sup_{t \in \mathcal{T}} \Big\| \sum_{I \in \mathcal{I}_{n,d}} X_I t_I \Big\|, \tag{4.6}$$

where $X_I := \prod_{i \in I} X_i$. For any $k \in [d]$ we let

$$\begin{aligned} W_k &:= \sup_{t \in \mathcal{T}} \sup_{v^* \in \mathcal{B}_1^*} \sup_{\substack{\alpha^1, \ldots, \alpha^k \in \mathbb{R}^n \\ |\alpha^i| \le 1}} v^* \Big(\sum_{\substack{i_1, \ldots, i_k \\ \text{distinct}}} \alpha^1_{i_1} \cdots \alpha^k_{i_k} \sum_{\substack{I \in \mathcal{I}_{n,d-k} \\ i_1, \ldots, i_k \notin I}} X_I t_{I \cup \{i_1, \ldots, i_k\}} \Big) \\ &= \sup_{t \in \mathcal{T}} \sup_{\substack{\alpha^1, \ldots, \alpha^k \in \mathbb{R}^n \\ |\alpha^i| \le 1}} \Big\| \sum_{\substack{i_1, \ldots, i_k \\ \text{distinct}}} \alpha^1_{i_1} \cdots \alpha^k_{i_k} \sum_{\substack{I \in \mathcal{I}_{n,d-k} \\ i_1, \ldots, i_k \notin I}} X_I t_{I \cup \{i_1, \ldots, i_k\}} \Big\|, \end{aligned} \tag{4.7}$$

where for $k = d$ we use the convention $\mathcal{I}_{n,0} = \{\emptyset\}$ and $X_\emptyset := 1$. One can interpret the quantities W_k in the following way: If $f_t(x) = \sum_{I \in \mathcal{I}_{n,d}} x_I t_I$ is the corresponding polynomial in n variables, and $\partial^k f_t(x)$ is the k-tensor of all partial derivatives of order k, then $W_k = \sup_{t \in \mathcal{T}} |\partial^k f_t(X)|_{\mathrm{op}}$.

Furthermore, the concentration inequalities are phrased with the help of the quantities

$$\begin{aligned} \widetilde{W}_k &:= \sup_{\substack{\alpha^1, \ldots, \alpha^k \in \mathbb{R}^n \\ |\alpha^i| \le 1}} \sum_{\substack{i_1, \ldots, i_k \\ \text{distinct}}} \alpha^1_{i_1} \cdots \alpha^k_{i_k} \sup_{t \in \mathcal{T}} \sup_{v^* \in \mathcal{B}_1^*} v^* \Big(\sum_{\substack{I \in \mathcal{I}_{n,d-k} \\ i_1, \ldots, i_k \notin I}} X_I t_{I \cup \{i_1, \ldots, i_k\}} \Big) \\ &= \sup_{\substack{\alpha^1, \ldots, \alpha^k \in \mathbb{R}^n \\ |\alpha^i| \le 1}} \sum_{\substack{i_1, \ldots, i_k \\ \text{distinct}}} \alpha^1_{i_1} \cdots \alpha^k_{i_k} \sup_{t \in \mathcal{T}} \Big\| \sum_{\substack{I \in \mathcal{I}_{n,d-k} \\ i_1, \ldots, i_k \notin I}} X_I t_{I \cup \{i_1, \ldots, i_k\}} \Big\|. \end{aligned}$$

Clearly $\widetilde{W}_k \ge W_k$ holds for all $k \in [d]$. Concentration properties for functions as in (4.6) have been studied in the case of Rademacher random variables and in the real case in [BBLM05, Theorem 14] for all $d \ge 2$, and under certain technical assumptions in [Ada15]. We prove deviation inequalities in the weakly dependent setting, and afterwards discuss how these compare to the particular result in [BBLM05].

Theorem 4.4. *Let $X = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{R}^n$ with support in $[a,b]^n$ satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. For $f = f(X)$ as in (4.6) and all $p \ge 2$ we have*

$$\|(f - \mathbb{E} f)_+\|_p \le \sum_{j=1}^d (\sigma^2 (b-a)^2 (p-1))^{j/2} \, \mathbb{E} W_j, \tag{4.8}$$

$$\|f - \mathbb{E} f\|_p \le \sum_{j=1}^d (\sigma^2 (b-a)^2 (p-1))^{j/2} \, \mathbb{E} \widetilde{W}_j. \tag{4.9}$$

Consequently, for any $t \ge 0$ the following inequalities hold:

$$\mathbb{P}(f - \mathbb{E} f \ge t) \le e \exp\Big(- \frac{1}{\sigma^2 (b-a)^2} \min_{k \in [d]} \Big(\frac{t}{d e \, \mathbb{E} W_k} \Big)^{2/k} \Big), \tag{4.10}$$

$$\mathbb{P}(|f - \mathbb{E} f| \ge t) \le e \exp\Big(- \frac{1}{\sigma^2 (b-a)^2} \min_{k \in [d]} \Big(\frac{t}{d e \, \mathbb{E} \widetilde{W}_k} \Big)^{2/k} \Big). \tag{4.11}$$

The upper bound (4.10) in the case of Rademacher random variables has been treated in [BBLM05] and yields similar results. For a compact set of vectors $\mathcal{T}$ in $\mathbb{R}^{\mathcal{I}_{n,d}}$, [BBLM05, Theorem 14, Corollary 4] provides deviation inequalities for

$$f = f(X) \coloneqq f_{\mathcal{T}}(X) \coloneqq \sup_{t \in \mathcal{T}} \Big| \sum_{I \in \mathcal{I}_{n,d}} X_I t_I \Big| \tag{4.12}$$

by showing that for all $p \geq 2$

$$\|(f - \mathbb{E}\, f)_+\|_p \leq \sum_{j=1}^{d} (4\kappa p)^{j/2}\, \mathbb{E}\, W_j,$$

where $\kappa \approx 1.27$ is a numerical constant (cf. Theorem 2.12) and W_k is the quantity from equation (4.7) in the special case of $\mathcal{B} = \mathbb{R}$. Theorem 4.4 is comparable to this result. On the one hand, it is less general for independent random variables, but on the other hand it is valid in any Banach space and without the independence assumption on $X_1, \ldots, X_n$. If we consider Rademacher random variables, the following corollary follows from Theorem 4.4.

Corollary 4.5. *Let $X_1, \ldots, X_n$ be independent Rademacher random variables and $f = f(X)$ as in* (4.12)*. We have for any $p \geq 2$*

$$\|(f - \mathbb{E}\, f)_+\|_p \leq \sum_{k=1}^{d} (4(p-1))^{k/2}\, \mathbb{E}\, W_k.$$

Consequently, for any $t \geq 0$

$$\mathbb{P}\,(f - \mathbb{E}\, f \geq t) \leq e \exp\Big(-\frac{1}{4} \min_{k \in [d]} \Big(\frac{t}{de\, \mathbb{E}\, W_j} \Big)^{2/k} \Big).$$

Apart from Rademacher random variables, possible applications of Theorem 4.4 include the weakly dependent spin systems. One such example is given by the Curie–Weiss model on n sites with inverse temperature $\beta < 1$, i. e. the probability measure on $\{-1, +1\}^n$ defined by

$$\mu(x) \coloneqq Z^{-1} \exp\Big(\beta n^{-1} \sum_{i \neq j} x_i x_j\Big) \text{ where } Z = \sum_{x \in \{-1,+1\}^n} \exp\Big(\beta n^{-1} \sum_{i \neq j} x_i x_j\Big).$$

It can be easily seen that μ satisfies the conditions of Proposition 2.15 and thus a $\mathfrak{d}-\mathrm{LSI}(\sigma^2(\beta))$ holds.

As a second corollary, Theorem 4.4 can be used to partly recover Talagrand's Theorem 1.3 on concentration properties of quadratic forms in Banach spaces, and extend it to weakly dependent random variables. Considering the case $d = 2$, we can write

$$T_1 \coloneqq \mathbb{E}\, W_1 = \mathbb{E} \sup_{t \in \mathcal{T}} \sup_{v^* \in \mathcal{B}_1^*} \Big(\sum_{i=1}^n \Big(\sum_{j=1}^n X_j v^*(t_{ij}) \Big)^2 \Big)^{1/2}$$

$$T_2 \coloneqq \mathbb{E}\, W_2 = \sup_{t \in \mathcal{T}} \sup_{v^* \in \mathcal{B}_1^*} \|(v^*(t_{ij}))_{i,j}\|_{\mathrm{op}}.$$

Corollary 4.6. *Assume that $X = (X_1, \dots, X_n)$ satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ and is supported in $[a,b]^n$ and let $f_{\mathcal{T}} = f_{\mathcal{T}}(X)$ be as in* (4.6) *with $d = 2$. We have for all $t \geq 0$*

$$\mathbb{P}\left(f_{\mathcal{T}}(X) - \mathbb{E}\, f_{\mathcal{T}}(X) \geq t\right) \leq e \exp\Big(- \frac{1}{30(b-a)^2\sigma^2} \min\Big(\frac{t^2}{T_1^2}, \frac{t}{T_2} \Big)\Big).$$

Finally, let us remark that the case of independent Rademacher random variables $X_1, \dots, X_n$, there is a connection between multilevel concentration and quantities arising Boolean analysis. Recall that any function $f : \{-1, +1\}^n \to \mathbb{R}$ can be decomposed using the orthonormal *Fourier–Walsh basis* given by $(x_S)_{S \subseteq [n]}$ for $x_S \coloneqq \prod_{i \in S} x_i$. More precisely, we have

$$f(x) = \sum_{S \subseteq [n]} \hat{f}_S x_S = \sum_{j \in [n]} \sum_{S \subseteq [n] : |S| = j} \hat{f}_S x_S,$$

where $(\hat{f}_S)_{S \subseteq [n]}$ are the *Fourier coefficients* of f given by $\hat{f}_S = \int X_S f(X) d\,\mathbb{P}$. For any $j \in [n]$ we define the *Fourier weight of order j* as $W_j(f) \coloneqq \sum_{S \subseteq [n] : |S| = j} \hat{f}_S^2$. It is clear that $\|f\|_2^2 = \sum_{j=0}^n W_j(f)$. The following multilevel concentration inequality can now be easily deduced.

Proposition 4.7. *Let $X_1, \dots, X_n$ be independent Rademacher random variables and let $f : \{1, +1\}^n \to \mathbb{R}$ be a function given in the Fourier–Walsh basis as $f(x) = \sum_{j=0}^d \sum_{S \subset [n] : |S| = j} \hat{f}_S x_S$ for some $d \in \mathbb{N}, d \leq n$. For any $t \geq 0$ we have*

$$\mathbb{P}(|f(X) - \mathbb{E}\, f(X)| \geq t) \leq e \exp\Big(- \min_{j \in [d]} \Big(\frac{t}{deW_j(f)^{1/2}} \Big)^{2/j} \Big).$$

In other words, the event $|f(X) - \mathbb{E}\, f(X)| \leq de \max_{j \in [d]} (W_j t^j)^{1/2}$ holds with probability at least $\exp(1-t)$.

4.2.2 Concentration properties of U-statistics

Next, we use Theorem 4.1 to prove concentration properties of so-called U-*statistics* which frequently arise in statistical theory. We refer to [PG99] for an excellent monograph. More recently, concentration inequalities for U-statistics have been considered in [Ada06], [AW15, Section 3.1.2] and [BGS19, Corollary 1.3].

Let $\mathcal{Y} = \mathcal{X}^n$ and assume that $X_1, \ldots, X_n$ are either independent random variables, or the vector $X = (X_1, \ldots, X_n)$ satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. Let $h : \mathcal{X}^d \to \mathbb{R}$ be a measurable, symmetric function with $h(X_{i_1}, \ldots, X_{i_d}) \in L^\infty(\mathbb{P})$ for any $i_1, \ldots, i_d$, and define $B \coloneqq \|h(X_{i_1}, \ldots, X_{i_d})\|_{L^\infty(\mathbb{P})}$. We are interested in the concentration properties of the U-statistic with kernel h, i. e. of the random variable

$$f(X) = \sum_{i_1 \neq \ldots \neq i_d} h(X_{i_1}, \ldots, X_{i_d}). \tag{4.13}$$

Proposition 4.8. *Let $X = (X_1, \ldots, X_n)$ be as above and $f = f(X)$ be as in* (4.13)*. There exists a constant $C > 0$ depending on d (and σ^2 in the weakly dependent case) such that for any $t \geq 0$ the inequalities*

$$\mathbb{P}\Big(|f - \mathbb{E} f| \geq Bt\Big) \leq 2 \exp\Big(- \frac{1}{4C_d} \min_{k \in [d]} \Big(\frac{t}{\binom{d}{k} n^{d-k/2}} \Big)^{2/k} \Big) \tag{4.14}$$

$$\mathbb{P}\Big(n^{1/2-d} |f - \mathbb{E} f| \geq Bt\Big) \leq 2 \exp\Big(- \frac{1}{4C_d} \min\Big(t^2, n^{1-1/d} t^{2/d}\Big)\Big) \tag{4.15}$$

hold.

The normalization $n^{1/2-d}$ in (4.15) is in good agreement for U-statistics generated by a symmetric, non-degenerate kernel h in independent random variables (where non-degeneracy means $\mathrm{Var}(\mathbb{E}_{X_1} h(X_1, \ldots, X_d)) > 0$). Indeed, in the i. i. d. case [PG99, Remarks 4.2.5] states that

$$n^{1/2-d} \sum_{i_1 \neq \ldots \neq i_d} h(X_{i_1}, \ldots, X_{i_d}) \Rightarrow \mathcal{N}(0, d^2 \,\mathrm{Var}(\mathbb{E}_{X_1} h(X_1, \ldots, X_d)))$$

whenever $\mathbb{E}\, h(X_{i_1}, \ldots, X_{i_d}) = 0$ and $\mathbb{E}\, h(X_1, \ldots, X_d)^2 < \infty$. Actually, (4.15) shows that for $t \leq \sqrt{n}$ we have sub-Gaussian tails for any finite $n \in \mathbb{N}$ for bounded kernels h, although the variance of the limit distribution might be grossly overestimated.

4.2.3 Polynomials in the Ising model

Yet another application of Theorem 4.1 and Proposition 4.3 are concentration results for homogeneous polynomials in spins in the Ising model with bounded coefficients, and suitably recentered versions of a d-th order chaos. To begin with, let us consider the case of a weakly dependent Ising model without external field.

Theorem 4.9. *Let μ be an Ising model as in Proposition 2.15 with $h = 0$, and $X \sim \mu$. Let $A = (a_{i_1, \ldots, i_d})$ be a symmetric d-tensor with vanishing diagonal and $\sup_{i_1, \ldots, i_d} |a_{i_1 \ldots i_d}| = 1$, and define $f : \{-1, +1\}^n \to \mathbb{R}$ via*

$$f \coloneqq f(\sigma) \coloneqq \sum_{I = (i_1, \ldots, i_d)} a_I \sigma_I.$$

There is a constant $c = c(d, \alpha) > 0$ such that for all $t \geq 0$ it holds

$$\mathbb{P}\Big(|f(X) - \mathbb{E}\, f(X)| \geq t\Big) \leq 2\exp\Big(-\frac{1}{cn}\min(t^2, t^{2/d})\Big). \tag{4.16}$$

This result improves upon [GLP18, Theorem 1] as well as on [DDK17, Theorem 5] by removing the logarithmic dependence in the exponential. More precisely, in [GLP18] it is shown that for some weakly dependent Ising models π without external field, every degree d polynomial f with coefficients in $[-K,K]$ satisfies

$$\pi\big(|f - \mathbb{E}_\pi f| \geq t\big) \leq Cn^{d^2}\exp\Big(-\frac{t^{2/d}}{CnK^{2/d}}\Big).$$

Similar concentration inequalities have been proven in [DDK17]; given a degree d multilinear polynomial f in the spin variables of an Ising model π (in an α-high temperature regime without external field, i. e. for models satisfying the weak dependence condition) with coefficients in $[-K,K]$, [DDK17, Theorem 5] states that there are two constants such that for any $t \geq CK(n\log^2 n)^{d/2}$ we have

$$\pi\big(|f - \mathbb{E}_\pi f| \geq t\big) \leq 2\exp\Big(-\frac{\alpha t^{2/d}}{CK^{2/d}n\log n}\Big).$$

In contrast, (4.16) is optimal in terms of the dependence on n and with respect to the power of t. To see this, note that the uniform measure $\mu = \otimes_{i=1}^n \frac{1}{2}(\delta_{-1} + \delta_{+1})$ can also be interpreted as an Ising model. In this case, for the tensor $A = (a_{i_1,\ldots,i_d})$ with entries $a_{i_1,\ldots,i_d} = 1$ if $i_1 \neq \ldots \neq i_d$ and 0, else, we have $\mathrm{Var}(f) \sim n^d$, so that f needs to be normalized by $n^{-d/2}$. Regarding the power of t, the invariance principle in [MOO10, Theorem 2.1] shows that for the same multilinear form f as above the distributions of $f(X)$ and $f(G)$ for a Gaussian vector G are close in Kolmogorov distance. On the other hand, the behavior of a Gaussian chaos is known (see e. g. [Lat06]). Consequently, the decay $t^{2/d}$ is the correct one for large values of t.

Note that in the case of Rademacher random variables (i. e. $h = 0, J = 0$) we have $\mathrm{Var}(f) = |A|_{\mathrm{HS}}^2$, so that a normalization of f by $|A|_{\mathrm{HS}}$ should be sufficient. On the other hand, Theorem 4.9 suggests a normalization by $n^{d/2}|A|_\infty$, which might of a different order.

We can invoke Proposition 4.3 to strengthen the result slightly. At the same time, we consider Ising models with external fields $h \neq 0$. Note that the major difference to the Ising model without an external field is the loss of spin symmetry, i. e. the map $\sigma \mapsto -\sigma$ does not preserve the measure μ (more precisely, the push-forward is an Ising model with external field $-h$). Thus, the odd degree polynomials $\sigma_{i_1}\cdots\sigma_{i_{2k+1}}$ are not centered.

The next corollary is an immediate consequence of Proposition 4.3.

Corollary 4.10. *Let μ be an Ising model as in Proposition 2.15, with an external field h and let $X \sim \mu$. Suppose that $d \in \mathbb{N}$ and $A = (A_{i_1,\ldots,i_d})_{i_1,\ldots,i_d}$ is a symmetric d-tensor with vanishing diagonal and $f_{d,A}$ as in* (4.3). *There exists a constant*

$C = C(\alpha,\widetilde{\alpha},d) > 0$ *such that for all* $t \geq 0$ *it holds*

$$\mathbb{P}\left(|f_{d,A}(X)| \geq t\right) \leq 2\exp\Big(-\frac{t^{2/d}}{C|A|_{\mathrm{HS}}^{2/d}}\Big). \tag{4.17}$$

Thus, (4.17) shows that $f_{d,A}$ needs to be normalized by the Hilbert–Schmidt norm of the d-tensor A. For $d \in \{1,2,3,4\}$ the functions $f_{d,A}$ are very explicit, since we have

$$\begin{aligned}
f_{1,A}(X) &= \sum_{i=1}^n a_i\widetilde{X_i}, \\
f_{2,A}(X) &= \sum_{i,j=1}^n a_{ij}(\widetilde{X}_{ij} - \mathbb{E}\,\widetilde{X}_{ij}), \\
f_{3,A}(X) &= \sum_{i,j,k=1}^n a_{ijk}\big(\widetilde{X}_{ijk} - \mathbb{E}\,\widetilde{X}_{ijk} - 3\widetilde{X}_i\,\mathbb{E}(\widetilde{X}_{jk})\big), \\
f_{4,A}(X) &= \sum_{i,j,k,l=1}^n a_{ijkl}\big(\widetilde{X}_{ijkl} - \mathbb{E}\,\widetilde{X}_{ijkl} - 4\widetilde{X_i}\,\mathbb{E}\,\widetilde{X}_{jkl} - 6\widetilde{X}_{ij}\,\mathbb{E}\,\widetilde{X}_{kl} + 6\,\mathbb{E}\,\widetilde{X}_{ij}\,\mathbb{E}\,\widetilde{X}_{kl}\big).
\end{aligned}$$

The $d = 3$ case has an interesting interpretation. Assume that $h = 0$, so that the X_i are centered random variables. Equation (4.17) states that a polynomial of order three is not concentrated around its mean (which in this case would be zero), but around a first order correction. For example, for the 3-tensor $A = (a_{ijk})$ given by $a_{ijk} = n^{-3/2}\mathbb{1}_{\Delta_3}(i,j,k)$, we obtain concentration inequalities for

$$f(X) = n^{-3/2}\sum_{i\neq j\neq k} X_{ijk} - 3n^{-3/2}\sum_{i=1}^n X_i \sum_{j\neq k: j\neq i, k\neq i} \mathrm{Cov}(X_j, X_k) =: f_3(X) + f_1(X).$$

Here, the correction term f_1 is sub-Gaussian, as a short calculation shows that we have $|\mathfrak{h}f_1|^2 = 2n^{-3}\sum_{i=1}^n c_i^2$ for $c_i \coloneqq 3\,\mathbb{E}\sum_{j\neq k: j\neq i, k\neq i} X_jX_k$, and [GLP18, Lemma 3.1] yields $c_i^2 \leq Cn^2$ for any $i \in \{1,\dots,n\}$.

4.2.4 Number of triangles in exponential random graph models

The last application will be the extension of a rather classical question in random graph theory to the setting of dependent edges. In the Erdös–Rényi model, the asymptotic properties of the number of triangles are quite classical and well-studied. For example, there are various conditions ensuring that the subgraph counts are asymptotically normal, see [Ruc88]. Therefore, it is an interesting task to find analogous results for the exponential random graph model.

Although the edges in this model are dependent, a weak dependence condition should suffice to expect a similar behavior as in the case of independent edges. For example, the large deviation results in [CD13] imply that in certain cases an exponential random graph model is indistinguishable from an Erdös–Rényi model in the limit. Although large deviation results are purely asymptotic, one can still hope for similar behavior concerning certain statistics for finite n.

Recall that the exponential random graph model μ_β is a spin system with sites $\mathcal{I}_n := \{(i,j) \in [n]^2 : i < j\}$. We let $\binom{\mathcal{I}_n}{3}$ be the set of all possibilities of choosing three distinct edges and

$$\mathcal{T}_n := \Big\{\{e,f,g\} \in \binom{\mathcal{I}_n}{3} : e,f,g \text{ form a triangle}\Big\}. \tag{4.18}$$

The statistic we are interested in is the number of triangles

$$T_3(x) := \sum_{\{e_1,e_2,e_3\}\in\mathcal{T}_n} x_{e_1}x_{e_2}x_{e_3}. \tag{4.19}$$

We prove multilevel concentration inequalities for T_3 and a linear approximation thereof. Define $\mu_2 := \mathbb{E}_{\mu_\beta} x_e x_f$ (for some edges $e \neq f \in \mathcal{I}_n$, $e \cap f \neq \emptyset$) and $f_1 := \sum_{e\in\mathcal{I}_n}(x_e - \mathbb{E}_{\mu_\beta}(x_e))$. From the definition of the ERGM it is clear that μ_2 is well-defined.

Theorem 4.11. *Let μ_β be an ERGM satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. There exists a constant $C = C(\sigma^2) > 0$ such that for all $t \geq 0$ we have the multilevel concentration bounds*

$$\mu_\beta(|T_3 - \mathbb{E}_{\mu_\beta} T_3| \geq t) \leq 2\exp\Big(-\frac{1}{C}\min\Big(\Big(\frac{t}{n^{3/2}}\Big)^{2/3}, \frac{t}{\mu_1 n^{3/2}}, \Big(\frac{t}{\mu_2 n^2}\Big)^2\Big)\Big) \tag{4.20}$$

$$\mu_\beta(|T_3 - \mathbb{E}_{\mu_\beta} T_3 - (n-2)\mu_2 f_1| \geq t) \leq 2\exp\Big(-\frac{1}{C}\min\Big(\Big(\frac{t}{n^{3/2}}\Big)^{2/3}, \frac{t}{\mu_1 n^{3/2}}\Big)\Big). \tag{4.21}$$

It is interesting to observe effect of subtracting the random variable $(n-2)\mu_2 f_1$. As the variance of T_3 is of order n^4, a normalization by n^{-2} is necessary to obtain a stable variance, and (4.20) gives suitable tail estimates. However, the random variable $T_3 - \mathbb{E}_{\mu_\beta} T_3 - (n-2)\mu_2 f_1$ concentrates on a narrower range, since the variance is of order n^3, and (4.21) yields stretched-exponential tails in this case. In the Erdös–Rényi model, a short calculation shows

$$\begin{aligned} \mathrm{Var}\,(T_3) &= \binom{n}{3}p^3(1-p^3) + \frac{1}{2}n(n-1)(n-2)(n-3)p^5(1-p) \\ \mathrm{Var}\,\big(T_3 - (n-2)p^2 f_1\big) &= \binom{n}{3}p^3(1-p^3). \end{aligned}$$

To complement these observations, inspecting (4.20), we see that the normalization n^{-2} corresponds to the factor n^{-4} in the Gaussian part, whereas the exponential and stretched-exponential part require a normalization by $n^{-3/2}$ only.

The inequality (4.21) shows that T_3 fluctuates around the linear term f_1 on a lower order. This leads to the idea of mimicking the method of Hájek projection

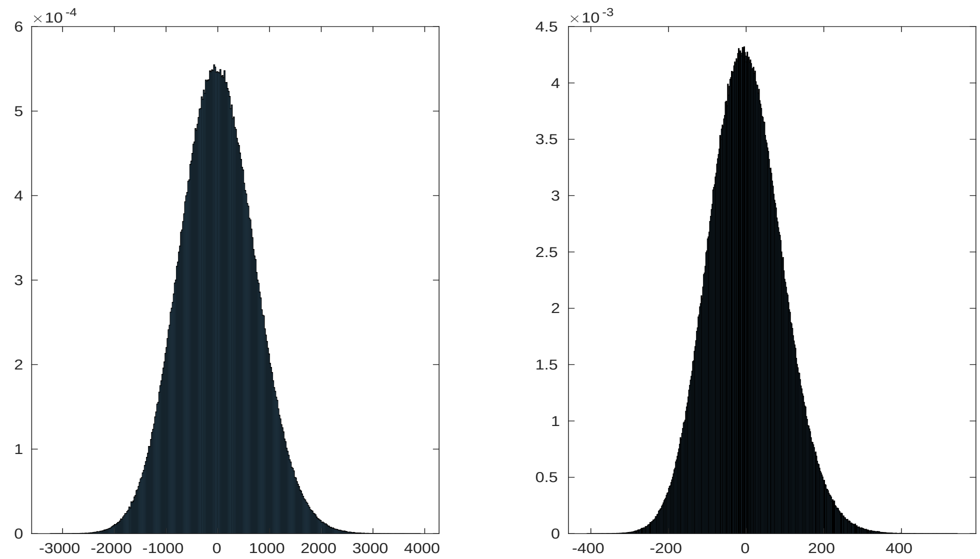

Figure 4.1: A comparison of the distributions of $T_3 - \mu_\beta(T_3)$ (left) and $T_3 - \mu_\beta(T_3) - (n-2)\mu_2 f_1$ (right) for $n = 100$, $\beta_1 = -0.1, \beta_2 = 0.05$ and $G_1 = K_2$ (an edge), $G_2 = K_3$ (a triangle) using the Glauber dynamics and roughly 2 million simulations.
The sample standard deviation of T_3 is approximately 725, whereas for the linear approximation it is 94.

to deduce a central limit theorem for the triangle count from a CLT for the edge count. As far as we are aware, there are hardly any theoretical results on the distributional limits of the subgraph counts as $n \to \infty$ barring certain special cases. (One such example is the edge two-star model, which can also be interpreted as an Ising model, see [Muk13a; Muk13b].)

Corollary 4.12. *Let μ_β be an ERGM satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. Assuming the central limit theorem $\binom{n}{2}^{-1/2} \sum_{e \in \mathcal{I}_n} (x_e - \mathbb{E}_{\mu_\beta} x_e) \Rightarrow \mathcal{N}(0,v^2)$, we can infer*

$$\frac{T_3 - \mu_\beta(T_3)}{(n-2)\mu_2 \sqrt{\binom{n}{2}}} \Rightarrow \mathcal{N}(0,v^2).$$

The formulation of Corollary 4.12 is slightly sloppy, as we actually consider the sequence of probability measures induced on the graphs on n vertices.

Furthermore, the convergence can be quantified in the Wasserstein distance. Let us recall that for two probability measures μ, ν on $\mathbb{R}$ with finite first moment (i. e. $\int |x| d\mu(x) < \infty, \int |x| d\nu(x) < \infty$) the Wasserstein distance is defined as

$$d_W(\mu,\nu) = \sup \Big\{ \int f d\mu - \int f d\nu : f \in \mathrm{Lip}_1 \Big\},$$

where Lip_1 denotes the set of all Lipschitz-continuous functions with Lipschitz constant at most 1. For two random variables X,Y we define $d_W(X,Y)$ as the Wasserstein distance between their distributions. Define

$$\begin{aligned}\widetilde{T}_3(x) &:= (n-2)^{-1}\mu_2^{-1}\binom{n}{2}^{-1/2}\sum_{\{e,f,g\}\in\mathcal{T}_n}(x_e x_f x_g - \mathbb{E}_{\mu_\beta}(x_e x_f x_g))\\ \widetilde{L}(x) &:= \binom{n}{2}^{-1/2}\sum_{e\in\mathcal{I}_n}(x_e - \mathbb{E}_{\mu_\beta} x_e).\end{aligned}$$

Proposition 4.13. *Let $\mu_\beta = \mu_\beta^{(n)}$ be an ERGM satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ and let $Z \sim \mathcal{N}(0,v^2)$ for some $v^2 > 0$. There exists a constant $C = C(\sigma^2)$ such that*

$$d_W(\widetilde{T}_3, Z) \le d_W(\widetilde{L}, Z) + Cn^{-1/2}.$$

Consequently, a rate of convergence in the Wasserstein distance for the number of edges immediately implies a rate of convergence for the number of triangles. As the number of edges is a linear function in the random variables $(x_e)_{e\in\mathcal{I}_n}$, this is a much easier object to handle than the third order polynomial T_3.

4.3 Proofs

In the next proofs, we need the inequality

$$\|W\|_p \le \mathbb{E}\,W + \|(W - \mathbb{E}\,W)_+\|_p \tag{4.22}$$

for any positive random variable W, which is an immediate consequence of the pointwise inequality $W \le (W - \mathbb{E}\,W)_+ + \mathbb{E}\,W$ and the Minkowski inequality.

Proof of Theorem 4.1. (1): Since $X_1, \ldots, X_n$ are independent, Theorem 2.12 and (4.22) yield

$$\|f - \mathbb{E}\,f\|_p \le (8\kappa p)^{1/2}\|\mathfrak{h}f\|_p \le (8\kappa p)^{1/2}\|\mathfrak{h}f\|_{\mathrm{op},1} + (8\kappa p)^{1/2}\|(|\mathfrak{h}f| - \mathbb{E}|\mathfrak{h}f|)_+\|_p.$$

The second term can be estimated using Theorem 2.12 again, which in combination with Lemma 2.6 gives

$$\|(|\mathfrak{h}f| - \mathbb{E}|\mathfrak{h}f|)_+\|_p \le \sqrt{2\kappa p}\|\mathfrak{h}^+|\mathfrak{h}f|\|_p \le \sqrt{2\kappa p}\|\mathfrak{h}^{(2)}f\|_{\mathrm{op},p}.$$

This can be iterated to obtain for any $d \in \mathbb{N}$

$$\|f - \mathbb{E}\,f\|_p \le \sum_{j=1}^{d-1}(8\kappa p)^{j/2}\|\mathfrak{h}^{(j)}f\|_{\mathrm{op},1} + (8\kappa p)^{d/2}\|\mathfrak{h}^{(d)}f\|_{\mathrm{op},\infty}.$$

Now it remains to apply Proposition 2.10.

(2): The proof of the second part is very similar. In the first step, using

Proposition 2.8 and (4.22) leads to

$$\|f - \mathbb{E} f\|_p \leq (\sigma^2(p-3/2))^{1/2}\|\mathfrak{h} f\|_{\mathrm{op},1} + (\sigma^2(p-3/2))^{1/2}\|(|\mathfrak{h} f| - \mathbb{E}|\mathfrak{h} f|)_+\|_p.$$

Equation (2.8) can be used to estimate the second term on the right hand side. So, for any $d \in \mathbb{N}$ we have by an iteration

$$\|f - \mathbb{E} f\|_p \leq \sum_{j=1}^{d-1} (\sigma^2(p-1))^{j/2}\|\mathfrak{h}^{(j)} f\|_{\mathrm{op},1} + (\sigma^2(p-1))^{d/2}\|\mathfrak{h}^{(d)} f\|_{\mathrm{op},\infty}$$

Again we can apply Proposition 2.10 to obtain the concentration inequality. □

To prove Proposition 4.3, it is convenient to introduce some notations. For any distinct indices $k_1, \ldots, k_s \in [d]$ and $l_1, \ldots, l_s \in \mathcal{I}$ let $A^{k_1=l_1,\ldots,k_s=l_s}$ be the $(d-s)$-tensor with fixed entries $k_i = l_i$ for all $i \in [s]$. For example, if $A = (A_{ijkl})$ is a 4-tensor, $A^{2=j,3=i}$ is the 2-tensor given by $A_{kl}^{2=j,3=i} = A_{kjil}$. Clearly, the symmetry and vanishing diagonal property are inherited.

Proof of Proposition 4.3. To see that $f_{d,A}$ has mean zero, fix d distinct indices $i_1, \ldots, i_d$ and an arbitrary partition $P \in \mathcal{P}(\{i_1, \ldots, i_d\})$. If $N(P) = 1$, then g_P has mean zero by construction, as the only stochastic term is a factor $\widetilde{f}_{i_j}$. Otherwise, for $N(P) \geq 2$, the partition is of the form $P = \{\{i_1\}, \ldots, \{i_{N(P)}\}, I_1, \ldots, I_l\}$, and $\widetilde{P} = \{\{i_1, \ldots, i_{N(P)}\}, I_1, \ldots, I_l\}$ is also a valid partition and $g_{\widetilde{P}} = \mathbb{E}\, g_P$. As a consequence, $\mathbb{E}\, f_{d,A} = 0$.

For any $l \in \mathcal{I}$ write T_l for the formal operator that replaces x_l by $\hat{x}_l$. First off, we have

$$\begin{aligned}
\mathfrak{h}_l f_{d,A}(X) &= \sup_{x_l,\hat{x}_l} \Big| \sum_I A_I \sum_{P \in \mathcal{P}(I)} (-1)^{M(P)} \left(g_P(X_I) - g_P(T_l(X_I))\right) \Big| \\
&= \sup_{x_l,\hat{x}_l} \Big| (f(x_l) - f(\hat{x}_l)) \sum_{k=1}^d \sum_{I=(i_1,\ldots,i_{d-1})} A_I^{k=l} \sum_{P \in \mathcal{P}(I)} (-1)^{M(P)} g_P(X_I) \Big| \\
&\leq c \Big| \sum_{k=1}^d \sum_{I=(i_1,\ldots,i_{d-1})} A_I^{k=l} \sum_{P \in \mathcal{P}(I)} (-1)^{M(P)} g_P(X_I) \Big| \\
&= c \Big| \sum_{k=1}^d f_{d-1,A^{k=l}} \Big|.
\end{aligned}$$

Here, the second equality follows from the fact that $T_l(x_{i_1}, \ldots, x_{i_d}) = (x_{i_1}, \ldots, x_{i_d})$ unless $i_j = l$ for some j and the definition of g_P, and the inequality in the third line is a consequence of the assumptions. We can assume $c = 1$, since the general case follows by rescaling f by c^{-1}. Now, by the $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ property it holds for any $p \geq 2$

$$\|f_{d,A}\|_p^2 \leq \|f_{d,A}\|_2^2 + \sigma^2(p-2)\|\mathfrak{h} f_{d,A}\|_p^2.$$

Using the Poincaré inequality (2.4) with respect to $\mathfrak{h}$ gives

$$\|f_{d,A}\|_2^2 \le \sigma^2 \sum_{l_1} \mathbb{E}(\mathfrak{h}_{l_1} f_{d,A})^2 \le \sigma^2 \sum_{l_1} \mathbb{E}(\widetilde{\mathfrak{h}}_{l_1} f_{d,A})^2 \le \sigma^2 \|\widetilde{\mathfrak{h}} f_{d,A}\|_p^2,$$

where $\widetilde{\mathfrak{h}}_l$ replaces $\sup_{x_l,\hat{x}_l} |f(x_l) - f(\hat{x}_l)|$ by 1. Clearly, since $\mathfrak{h}_l f_{d,A} \le \widetilde{\mathfrak{h}}_l f_{d,A}$ pointwise, the L^p-norms can be estimated as well, resulting in $\|f_{d,A}\|_p^2 \le \sigma^2 (p-1) \|\widetilde{\mathfrak{h}} f_{d,A}\|_p^2$. We have

$$\widetilde{\mathfrak{h}}_{l_1} f_{d,A} = \Big| \sum_{k_1=1}^d \sum_{I=(i_1,\dots,i_{d-1})} A_I^{k_1=l_1} \sum_{P \in \mathcal{P}(I)} (-1)^{M(P)} g_P \Big|,$$

which itself is the absolute value of a sum of centered random variables, so that the process can be iterated; in each step, the Poincaré inequality (2.4) can be used and

$$\widetilde{\mathfrak{h}}_{l_1} \cdots \widetilde{\mathfrak{h}}_{l_s} f_{d,A} = \Big| \sum_{k_1=1}^d \cdots \sum_{k_s=1}^{d-s} \sum_{I=(i_1,\dots,i_{d-s})} A_I^{k_1=l_1,\dots,k_s=l_s} \sum_{P \in \mathcal{P}(I)} (-1)^{M(P)} g_P \Big|.$$

Thus, using the inequality $|\widetilde{\mathfrak{h}}|\widetilde{\mathfrak{h}}^{(d)} f|| \le |\widetilde{\mathfrak{h}}^{(d+1)} f|$ and taking the square root yields

$$\|f_{d,A}\|_p \le \sigma^d |A| p^{d/2}$$

The concentration inequality follows from Proposition 2.10. □

Proof of Theorem 4.4. Let us first consider the case that X satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. Recall that we have by (2.8)

$$\|(f - \mathbb{E} f)_+\|_p \le (\sigma^2(p-1))^{1/2} \|\mathfrak{h}^+ f\|_p.$$

Next, we prove the inequality $|\mathfrak{h}^+ f| \le (b-a) W_1$. To see this, let $(\widetilde{t}, \widetilde{v}^*)$ be the tuple satisfying $\sup_{t \in \mathcal{T}} \sup_{v^* \in \mathcal{B}_1^*} v^*(\sum_{I \in \mathcal{I}_{n,d}} X_I t_I) = \widetilde{v}^*(\sum_{I \in \mathcal{I}_{n,d}} X_I \widetilde{t}_I)$, and observe that

$$\begin{aligned}
|\mathfrak{h}^+ f(X)|^2 &= \sum_{i=1}^n \sup_{x_i'} \Big(\sup_{t,v^*} v^*\Big(\sum_{I \in \mathcal{I}_{n,d}} X_I t_I \Big) - \sup_{t,v^*} v^*\Big(\sum_{I \in \mathcal{I}_{n,d}} (\overline{X}_i, x_i')_I t_I \Big) \Big)_+^2 \\
&\le \sum_{i=1}^n \sup_{x_i'} \Big((X_i - x_i') \sum_{I \in \mathcal{I}_{n,d-1} : i \notin I} \widetilde{v}^*(X_I \widetilde{t}_{I \cup \{i\}}) \Big)^2 \\
&\le (b-a)^2 \sum_{i=1}^n \widetilde{v}^*\Big(\sum_{I \in \mathcal{I}_{n,d-1} : i \notin I} X_I \widetilde{t}_{I \cup \{i\}} \Big)^2 \\
&= (b-a)^2 \sup_{\alpha^1 : |\alpha^1| \le 1} \widetilde{v}^*\Big(\sum_{i=1}^n \alpha_i^1 \sum_{I \in \mathcal{I}_{n,d-1} : i \notin I} X_I \widetilde{t}_{I \cup \{i\}} \Big)^2
\end{aligned}$$

$$\begin{aligned}
&\leq (b-a)^2 \Big(\sup_{t,v^*} \sup_{\alpha^1 : |\alpha^1| \leq 1} v^* \Big(\sum_{i=1}^n \alpha_i^1 \sum_{I \in \mathcal{I}_{n,d} : i \notin I} X_I t_{I \cup \{i\}} \Big) \Big)^2 \\
&= (b-a)^2 W_1^2.
\end{aligned}$$

Consequently,

$$\|(f - \mathbb{E}\, f)_+\|_p \leq (\sigma^2 (b-a)^2 (p-1))^{1/2} \left(\mathbb{E}\, W_1 + \|(W_1 - \mathbb{E}\, W_1)_+\|_p \right).$$

As in [BBLM05], this can now be iterated, i. e. we have for any $k \in [d-1]$ $|\mathfrak{h}^+ W_k| \leq (b-a) W_{k+1}$. Here we may argue as above, where the only difference is to choose $(\widetilde{t}, \widetilde{v}^*)$ and $\widetilde{\alpha}^1, \ldots, \widetilde{\alpha}^k$ which maximize W_k. This finally leads to

$$\|(f - \mathbb{E}\, f)_+\|_p \leq \sum_{j=1}^d (\sigma^2 (b-a)^2 (p-1))^{j/2}\, \mathbb{E}\, W_j,$$

using that W_d is constant. This proves (4.8).

Secondly, to prove (4.9), let us first consider why we cannot argue as before. Note that the argument heavily relies on the positive part of the difference operator $\mathfrak{h}^+$, which allows us to choose the maximizers $\widetilde{t}, \widetilde{v}$ independent of $i \in [n]$. This is no longer possible for the concentration inequality. Here, Proposition 2.8 yields

$$\begin{aligned}
\|f - \mathbb{E}\, f\|_p &\leq (\sigma^2 (p - 3/2))^{1/2} \|\mathfrak{h} f\|_p \\
\|(f - \mathbb{E}\, f)_+\|_p &\leq (\sigma^2 (p-1))^{1/2} \|\mathfrak{h}^+ f\|_p.
\end{aligned}$$

Thus this argument fails in the first step if we try to use these inequalities. However, we can rewrite $\mathfrak{h}_i f(x) = \sup_{x_i', x_i''} (f(\overline{x}_i, x_i') - f(\overline{x}_i, x_i''))_+ = \sup_{x_i'} \mathfrak{h}_i^+ f(\overline{x}_i, x_i')$, where the sup is to be understood with respect to the support of X_i'. As a consequence, we have for each fixed $i \in [n]$ (again choosing $\widetilde{t}$ by maximizing the first summand in the brackets)

$$\begin{aligned}
\mathfrak{h}_i f(x)^2 &= \sup_{x_i'} \sup_{x_i''} \Big(\sup_{t \in \mathcal{T}} \Big\| \sum_{I \in \mathcal{I}_{n,d}} (\overline{X}_i, x_i')_I t_I \Big\| - \sup_{t \in \mathcal{T}} \Big\| \sum_{I \in \mathcal{I}_{n,d}} (\overline{X}_i, x_i'')_I t_I \Big\| \Big)_+^2 \\
&\leq \sup_{x_i'} \sup_{x_i''} \Big\| (x_i' - x_i'') \sum_{I \in \mathcal{I}_{n,d-1} : i \notin I} X_I \widetilde{t}_{I \cup \{i\}} \Big\|^2 \\
&\leq (b-a)^2 \sup_{t \in \mathcal{T}} \Big\| \sum_{I \in \mathcal{I}_{n,d-1} : i \notin I} X_I t_{I \cup \{i\}} \Big\|^2.
\end{aligned}$$

This implies

$$|\mathfrak{h} f| \leq (b-a) \sup_{\substack{\alpha^1 \in \mathbb{R}^n \\ |\alpha^1| \leq 1}} \sum_{i=1}^n \alpha_i^1 \sup_{t \in \mathcal{T}} \Big\| \sum_{\substack{I \in \mathcal{I}_{n,d-1} \\ i \notin I}} X_I t_{I \cup \{i\}} \Big\| = (b-a) \widetilde{W}_1.$$

The proof is now easily completed as in the first part, with W_k replaced by $\widetilde{W}_k$. $\square$

Proof of Corollary 4.5. Since the uniform distribution on $\{-1,+1\}^n$ satisfies a $\mathfrak{d}-$LSI with constant 1 (see e. g. [Gro75, Theorem 3] or [DS96, Example 3.1]), this follows immediately from Theorem 4.4. $\square$

Proof of Proposition 4.7. The proposition can be proven using a similar technique as before, since the Hilbert–Schmidt norms of higher order difference act as Fourier projections. We choose to take an alternate route as follows. The proof of [ODo14, Theorem 9.21] shows that for any f with degree at most d and any $p \geq 2$ it holds

$$\|f(X) - \mathbb{E}\, f(X)\|_p \leq \sum_{j=1}^{d} (p-1)^{j/2} W_j(f)^{1/2}. \tag{4.23}$$

Thus the assertion follows from Proposition 2.10. $\square$

Proof of Proposition 4.8. Without loss of generality assume $B = 1$, as otherwise we can replace h by hB^{-1}. We apply Theorem 4.1 in the respective cases and make use of the general bound $\|\mathfrak{h}^{(k)} f\|_{\mathrm{op},1} \leq \|\mathfrak{h}^{(k)} f\|_{\mathrm{HS},\infty}$. For any distinct $j_1, \ldots, j_k$ write $\|\cdot\| = \|\cdot\|_{j_1,\ldots,j_k,\infty}$, so that

$$\begin{aligned}
\mathfrak{h}_{j_1\ldots,j_k} f &= \Big\| f + \sum_{l=1}^{k} (-1)^l \sum_{1 \leq s_1 < \ldots < s_l \leq k} T_{j_{s_1} \ldots j_{s_l}} f \Big\| \\
&= \Big\| \sum_{i_1 \neq \ldots \neq i_d} h(X_{i_1}, \ldots, X_{i_d}) + \sum_{l=1}^{k} (-1)^l \sum_{s_1 < \ldots < s_l} T_{j_{s_1} \ldots j_{s_l}} h(X_{i_1}, \ldots, X_{i_d}) \Big\| \\
&=: \Big\| \sum_{i_1 \neq \ldots \neq i_d} S_{i_1,\ldots,i_d}(h,X) \Big\|.
\end{aligned}$$

Now it is easy to see that $S_{i_1,\ldots,i_d}(h,X) = 0$ unless $\{j_1, \ldots, j_k\} \subset \{i_1, \ldots, i_d\}$ (for example, this follows if one writes the sum inside of the norm as $\prod_{i=1}^k (\mathrm{Id} - T_{j_i}) f$), and in these cases one can upper bound the supremum by $2^k B$, from which we infer

$$\mathfrak{h}_{j_1\ldots,j_k} f \leq \binom{d}{k} 2^k (n-k) \cdots (n-d+1) \leq \binom{d}{k} 2^k n^{d-k}.$$

Consequently, this leads to

$$\|\mathfrak{h}^{(k)} f\|_{\mathrm{HS},\infty} \leq \binom{d}{k} 2^k n^{d-k} n^{k/2} = \binom{d}{k} 2^k n^{d-k/2}.$$

Thus, Theorem 4.1 yields for some $C = C(d,\sigma^2)$ and any $t \geq 0$

$$\mathbb{P}\Big(|f - \mathbb{E}\, f| \geq t\Big) \leq 2 \exp\Big(-\frac{1}{C} \min_{k \in [d]} \Big(\frac{t}{\binom{d}{k} 2^k n^{d-k/2}} \Big)^{2/k} \Big).$$

For the second part, choose $n^{d-1/2}t$ for $t \geq 0$ to obtain

$$\mathbb{P}\left(n^{1/2-d}|f - \mathbb{E}\, f| \geq t\right) \leq 2\exp\Big(-\frac{1}{4C_d}\min_{k\in[d]} n^{\frac{k-1}{k}} t^{2/k}\Big).$$

A short calculation shows that the minimum is attained at $k = 1$ in the range $t \leq \sqrt{n}$ and at $k = d$ otherwise, i. e.

$$\mathbb{P}\left(n^{1/2-d}|f - \mathbb{E}\, f| \geq t\right) \leq 2\exp\Big(-\frac{1}{4C_d}\min(t^2, n^{1-1/d}t^{2/d})\Big). \tag{4.24}$$

□

Proof of Theorem 4.9. Theorem 4.1 implies for $f = f(X)$ the multilevel concentration inequality

$$\mathbb{P}(|f-\mathbb{E}\, f| \geq t) \leq 2\exp\Big(-\frac{1}{C_d\sigma^2}\min_{k\in[d-1]}\Big(\frac{t}{\|\mathfrak{h}^{(k)}f\|_{\mathrm{HS},2}}\Big)^{2/k} \wedge \frac{t^{2/d}}{\|\mathfrak{h}^{(d)}f\|_{\mathrm{HS},\infty}^{2/d}}\Big). \tag{4.25}$$

Now for any $k \in [d-1]$ we have

$$(\mathfrak{h}_{j_1,\ldots,j_k}f)^2 \leq 2^{2k}\binom{d}{k}^2\Big(\sum_{\substack{i_1,\ldots,i_{d-k}\\ i_1,\ldots,i_{d-k}\notin\{j_1,\ldots,j_k\}}} a_{i_1,\ldots,i_{d-k},j_1,\ldots,j_k}\sigma_{i_1}\cdots\sigma_{i_{d-k}}\Big)^2.$$

Thus, ignoring the constants depending on d and k, [GLP18, Lemma 3.1] gives

$$\|\mathfrak{h}^{(k)}f\|_{\mathrm{HS},2} = \Big(\sum_{j_1,\ldots,j_k}\|\mathfrak{h}_{j_1,\ldots,j_k}f\|_2^2\Big)^{1/2} \leq c_{k,d}n^{k/2},$$

as for each fixed $j_1, \ldots, j_k$ we integrate a polynomial of degree at most $2(d-k)$ with coefficients bounded by 1.

For $k = d$ we have $\|\mathfrak{h}^{(k)}f\|_\infty \leq c_d n^{d/2}$ as well, since $\mathfrak{h}_{j_1\ldots j_d}f(X) \leq 2^d$. Consequently, plugging in the estimates into (4.25) yields the claim. □

To prove Theorem 4.11, we apply the general result on the concentration of the polynomials $f_{d,A}$ (see (4.3)) with the spin function $f(x) = x$. Before we do so, let us give a simple example which already demonstrates some of the arguments we will use.

Example. Let μ_β be an ERGM satisfying a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$, and let $T_1(x) \coloneqq \sum_{e\in\mathcal{I}_n} x_e$ be the number of edges. We have for any $t \geq 0$

$$\mu_\beta(|T_1 - \mathbb{E}_{\mu_\beta} T_1| \geq t) \leq 2\exp\Big(-\frac{t^2}{11\sigma^2 n(n-1)}\Big). \tag{4.26}$$

In particular, setting $\eta \coloneqq \mathbb{E}_{\mu_\beta}(x_e)$ for an arbitrary$e \in \mathcal{I}_n$, we obtain a strong law of large numbers, i. e. $T_1/|\mathcal{I}_n| \to \eta$ a.s.

Moreover, for any two disjoint subsets $S_1, S_2 \subset [n]$, write $C(S_1,S_2) := \{e = (i,j) \in \mathcal{I}_n : \{i,j\} \cap S_1 \neq \emptyset, \{i,j\} \cap S_2 \neq \emptyset\}$ and let $T_{S_1,S_2} := \sum_{e \in \mathcal{I}_n} \mathbb{1}_{C(S_1,S_2)}(e) x_e$ be the number of edges between S_1 and S_2. For any $t \geq 0$ it holds

$$\mu_\beta(|T_{S_1,S_2} - \mathbb{E}_{\mu_\beta}(T_{S_1,S_2})| \geq t) \leq 2 \exp\Big(- \frac{t^2}{22\sigma^2 |S_1||S_2|} \Big). \tag{4.27}$$

Proof. Noting that $T_1 - \mathbb{E}_{\mu_\beta}(T_1) = f_{1,A}$ for $A = (1, \ldots, 1)$, (4.26) readily follows from Proposition 4.3.

To show (4.27), write $B = (B_e)_e := (\mathbb{1}_{C(S_1,S_2)}(e))_e$ to see that we have $T_{S_1,S_2} - \mathbb{E}_{\mu_\beta}(T_{S_1,S_2}) = \sum_{e \in \mathcal{I}_n} B_e \widetilde{x}_e = f_{1,B}$. Noting that $|C(S_1,S_2)| = |S_1||S_2|$ leads to $|B|^2 = \sum_{e \in \mathcal{I}_n} B_e^2 = |S_1||S_2|$. Applying Proposition 4.3 then yields (4.27).

To prove the strong law of large numbers, first note that by the intrinsic symmetry (i. e. a relabeling of the vertices $[n]$ and a respective relabeling of the edges will result in the same probability law), it is easy to see that $\mathbb{E}_{\mu_\beta} x_e$ does not depend on $e \in \mathcal{I}_n$. Thus, η is well-defined and $\mathbb{E}_{\mu_\beta}(T_1) = |\mathcal{I}_n|\eta$. Now, (4.26) yields $T_1/|\mathcal{I}_n| \to \eta$ in probability, and the rate of convergence is of order $\exp(-\Omega(n^2))$, which in turn implies convergence almost surely by the Borell-Cantelli lemma. □

In a similar vein, we may now prove Theorem 4.11.

Proof of Theorem 4.11. We shall express the number of triangles as a linear combination of polynomials of the type $f_{d,A}$. For the proof fix $n \in \mathbb{N}$ and let $X \sim \mu_\beta$. Moreover, for any triangle (in K_n) $\{e_1, e_2, e_3\}$ we define $\mu_\Delta := \mathbb{E}\, X_{e_1 e_2 e_3}$ and $\widetilde{\mu}_\Delta := \mathbb{E}\, \widetilde{X}_{e_1 e_2 e_3}$, for two neighboring edges $e_1 \neq e_2$ we let

$$\mu_2 := \mathbb{E}\, X_{e_1 e_2}, \widetilde{\mu}_2 := \mathbb{E}\, \widetilde{X}_{e_1 e_2}$$

and $\mu_1 := \mathbb{E}\, X_e$. Finally, $(e_1, e_2, e_3) \in \mathcal{T}_n$ shall indicate that $\{e_1, e_2, e_3\} \in \mathcal{T}_n$ and $e_1 < e_2 < e_3$ (with some fixed partial ordering of the edges).

Now it is not hard to verify that we have the decomposition

$$T_3(X) - \mathbb{E}\, T_3(X) = f_3(X) + \mu_1 f_2(X) + (n-2)\mu_2 f_1(X), \tag{4.28}$$

using the auxiliary functions

$$f_1(X) := \sum_{e \in \mathcal{I}_n} \widetilde{X}_e, \qquad f_2(X) := \sum_{\substack{e_1 < e_2 \\ e_1 \cap e_2 \neq \emptyset}} \Big(\widetilde{X}_{e_1 e_2} - \widetilde{\mu}_{e_1 e_2} \Big),$$

$$f_3(X) := \sum_{(e_1, e_2, e_3) \in \mathcal{T}_n} \Big(\widetilde{X}_{e_1 e_2 e_3} - \widetilde{\mu}_\Delta - \widetilde{X}_{e_1} \widetilde{\mu}_2 - \widetilde{X}_{e_2} \widetilde{\mu}_2 - \widetilde{X}_{e_3} \widetilde{\mu}_2 \Big).$$

Hence, after symmetrization of the sum, the triangle count is the sum of three terms $f_{d,A}$ for different tensors A_3, A_2, A_1. More precisely, we have $(A_3)_{efg} = \frac{1}{6} \cdot \mathbb{1}_{\{e,f,g\} \in \mathcal{T}_n}$, $(A_2)_{ef} = \frac{\mu_1}{2} \mathbb{1}_{e \cap f \neq \emptyset}$ and $(A_1)_e = (n-2)\mu_2$. An easy counting argument shows $|A_3| \sim n^{3/2}/6$, $|A_2| \sim \mu_1 n^{3/2}/2$ and $|A_1| \sim \mu_2 n^2/\sqrt{2}$.

Finally, an application of Proposition 4.3 yields for $T_3 = T_3(X)$

$$\|T_3 - \mathbb{E}\, T_3\|_p \leq (\sigma^2 |A_3|^{2/3} p)^{3/2} + (\sigma^2 |A_2| p) + (\sigma^2 |A_1|^2 p)^{1/2},$$
$$\|T_3 - \mathbb{E}\, T_3 - (n-2)\mu_2 f_1\|_p \leq (\sigma^2 |A_3|^{2/3} p)^{3/2} + (\sigma^2 |A_2| p).$$

As always, the concentration inequalities now follow from Proposition 2.10. □

Proof of Corollary 4.12. Theorem 4.11 can be used to show that for any $t \geq 0$

$$\mu_\beta\Big(\Big|\frac{T_3 - \mathbb{E}_{\mu_\beta} T_3 - (n-2)\mu_2 f_1}{(n-2)\mu_2 \sqrt{\binom{n}{2}}}\Big| \geq t\Big) \to 0 \text{ for } n \to \infty, \tag{4.29}$$

and thus by [Bil68, Theorem 3.1] and the assumption

$$\frac{T_3 - \mathbb{E}_{\mu_\beta} T_3}{(n-2)\mu_2 \sqrt{\binom{n}{2}}} = \frac{T_3 - \mathbb{E}_{\mu_\beta} T_3 - (n-2)\mu_2 f_1}{(n-2)\mu_2 \sqrt{\binom{n}{2}}} + \frac{1}{\sqrt{\binom{n}{2}}} f_1 \Rightarrow \mathcal{N}(0,\sigma^2).$$

□

Remark. Actually equation (4.29) can be quantified; by (4.21), the rate of convergence is of order $\exp(-\Omega(n^{1/3}))$, which also implies almost sure convergence.

Proof of Proposition 4.13. By the triangle inequality for d_W it suffices to prove $d_W(\widetilde{T}_3, \widetilde{L}) \leq Cn^{-1/2}$ for some constant C depending on σ^2. For any 1-Lipschitz function we have due to $|f(x) - f(y)| \leq |x - y|$, Theorem 4.11 and a change of variables $s = \mu_2 n^{1/2} t$

$$\begin{aligned}
\big|\, \mathbb{E}_{\mu_\beta} f(\widetilde{T}_3) - \mathbb{E}_{\mu_\beta} f(\widetilde{L})\big| &\leq \mathbb{E}_{\mu_\beta} |\widetilde{T}_3 - \widetilde{L}| = \int_0^\infty \mu_\beta(|\widetilde{T}_3 - \widetilde{L}| \geq t) dt \\
&\leq \int_0^\infty \mu_\beta\Big(|T_3 - (n-2)\mu_2 f_1| \geq (n-2)\mu_2 \binom{n}{2}^{1/2} t\Big) dt \\
&\leq 2 \int_0^\infty \exp\Big(-\frac{1}{C} \min\Big((\mu_2 n^{1/2} t)^{2/3}, \mu_2 n^{1/2} t\Big)\Big) dt \\
&\leq 2n^{-1/2} \mu_2^{-1} \int_0^\infty \exp\Big(-\frac{1}{C} \min(s^{2/3}, s)\Big) ds.
\end{aligned}$$

Taking the supremum over all $f \in \mathrm{Lip}_1$ finishes the proof. □

CHAPTER 5

Concentration inequalities for polynomials in independent random variables

In this chapter we study concentration properties of polynomials in independent random variables $X_1, \ldots, X_n$. It differs from Chapter 4 by allowing them to be unbounded, but its applicability is reduced in two ways: We cannot recover concentration inequalities for weakly dependent random variables, and we only consider polynomials instead of arbitrary functions. We use the notation $\|A\|_{\mathrm{HS}}$ for the Hilbert–Schmidt norm of a d-tensor A, $\|A\|_{\mathrm{op}}$ for its operator norm as defined in (2.1) and $\|A\|_\infty = \max_{i_1,\ldots,i_d} |a_{i_1 \ldots i_d}|$ for the supremum norm. Moreover, we let $\|x\|_2$ be the Euclidean norm of a vector $x \in \mathbb{R}^n$.

Compared to the Hanson–Wright inequality Theorem 1.2 presented in Chapter 1, our aim is to weaken the hypothesis of sub-Gaussianity. We consider independent random variables $X_1, \ldots, X_n$ which have an *α-sub-exponential tail decay* in the sense that there exist two constants c, C and a parameter $\alpha > 0$ such that for all $i \in [n]$ and $t \geq 0$

$$\mathbb{P}(|X_i| \geq t) \leq c \exp\Big(- \frac{t^\alpha}{C} \Big). \tag{5.1}$$

There are many interesting choices of X_i of this type, such as bounded random variables (for any $\alpha > 0$), random variables with a sub-Gaussian (for $\alpha = 2$) or sub-exponential distribution ($\alpha = 1$), or “fatter” tails such as Weibull random variables with shape parameter $\alpha \in (0,1]$.

Another suitable name for random variables satisfying (5.1) is *sub-Weibull*(α), since if Y has a symmetrized Weibull distribution with shape parameter α and scale parameter 1 (i. e. has Lebesgue-density $f_\alpha(x) = \alpha/2 |x|^\alpha \exp(-|x|^\alpha)$) we have

$$\mathbb{P}(|Y| \geq t) = \exp(-t^\alpha).$$

This terminology has been used in [KC18, Definition 2.2].

We reformulate condition (5.1) in terms of so-called *exponential Orlicz norms*, but we emphasize that these two concepts are equivalent. For any random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\alpha > 0$ define the (quasi-)norm

$$\|X\|_{\psi_\alpha} := \inf \Big\{ t > 0 \colon \mathbb{E} \exp\Big(\frac{|X|^\alpha}{t^\alpha} \Big) \leq 2 \Big\}, \tag{5.2}$$

adhering to the standard definition $\inf \emptyset = \infty$. Strictly speaking, this is a norm for $\alpha \geq 1$ only, since otherwise the triangle inequality does not hold. Nevertheless, the above expression makes sense for any $\alpha > 0$, and we choose to call it a norm in these cases as well. For some properties of the Orlicz norms in the case $\alpha \in (0,1]$ we refer to Appendix B. First we concentrate on values $\alpha = 2/q$ for some $q \in \mathbb{N}$, and later prove results for $\alpha \in (0,1]$. For illustration, we start with a simplified version of some of our results which may already be sufficient for certain applications. The first result is a concentration inequality which can be considered as a generalization of the Hanson–Wright inequality (1.8) to quadratic forms in random variables with α-sub-exponential tail decay.

Proposition 5.1. *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbb{E}\, X_i = 0, \mathbb{E}\, X_i^2 = \sigma_i^2$, $\|X_i\|_{\Psi_\alpha} \leq M$ for some $\alpha \in (0,1] \cup \{2\}$, and $A = (a_{ij})$ be a symmetric $n \times n$ matrix. There exists a constant $C = C(\alpha) > 0$ such that for any $t \geq 0$ we have*

$$\mathbb{P}\big(\big|\sum_{i,j} a_{ij} X_i X_j - \sum_{i=1}^n \sigma_i^2 a_{ii}\big| \geq t\big) \leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{t^2}{M^4\|A\|_{\mathrm{HS}}^2}, \Big(\frac{t}{M^2\|A\|_{\mathrm{op}}}\Big)^{\frac{\alpha}{2}}\Big)\Big).$$

The symmetry assumption can clearly be removed. Indeed, if $A = (a_{ij})$ is not symmetric, we can apply Proposition 5.1 to its additive symmetrization $\widetilde{A} = \frac{1}{2}(A + A^T)$. Note that $\sum_{i,j} a_{ij} X_i X_j = \sum_{i,j} \widetilde{a}_{ij} X_i X_j$, $\|\widetilde{A}\|_{\mathrm{HS}}^2 \leq \|A\|_{\mathrm{HS}}^2$ and $\|\widetilde{A}\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}}$. As we will see in Proposition 5.5, the tail decay $\exp(-t^{\alpha/2}\|A\|_{\mathrm{op}}^{-\alpha/2})$ (for large t) can be sharpened by replacing the operator norm by a smaller norm. Actually, the technical result contains up to four different regimes instead of two as above.

The next theorem provides tail estimates for general polynomials. Note that this is not a generalization of Proposition 5.1 due to the use of the Hilbert–Schmidt instead of the operator norms.

Theorem 5.2. *Let $X_1, \ldots, X_n$ be independent random variables with $\|X_i\|_{\Psi_\alpha} \leq M$ for some $\alpha \in (0,1] \cup \{2\}$ and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree $D \in \mathbb{N}$. Then for all $t \geq 0$ it holds*

$$\mathbb{P}(|f(X) - \mathbb{E} f(X)| \geq t) \leq 2\exp\Big(-\frac{1}{C_{D,\alpha}}\min_{1 \leq d \leq D}\Big(\frac{t}{M^d\|\mathbb{E} f^{(d)}(X)\|_{\mathrm{HS}}}\Big)^{\frac{\alpha}{d}}\Big).$$

In particular, if $\|\mathbb{E} f^{(d)}(X)\|_{\mathrm{HS}} \leq 1$ for $d = 1, \ldots, D$, then

$$\mathbb{E}\exp\Big(\frac{C_{D,\alpha}}{M^\alpha}|f(X) - \mathbb{E}\, f(X)|^{\frac{\alpha}{D}}\Big) \leq 2,$$

or equivalently

$$\|f(X) - \mathbb{E}\, f(X)\|_{\Psi_{\frac{\alpha}{D}}} \leq C_{d,\alpha} M^D.$$

Informally, Theorem 5.2 states that a polynomial in random variables with tail decay as in (5.1) also exhibits α-sub-exponential tail decay whenever the Hilbert–

Schmidt norms are not too large. Moreover, the tail decay is "as expected", i. e. one just needs to account for the total degree D by taking the D-th root.

In particular, we can consider d-th order chaoses. That is, given a symmetric d-tensor $A = (a_{i_1 \dots i_d})$, we let

$$f_{d,A}(X) \coloneqq \sum_{i_1,\dots,i_d} a_{i_1\dots i_d}(X_{i_1} - \mathbb{E}X_{i_1}) \cdots (X_{i_d} - \mathbb{E}X_{i_d}). \tag{5.3}$$

In this situation, Theorem 5.2 reads as follows:

Corollary 5.3. *Let $X_1, \dots, X_n$ be independent random variables with $\|X_i\|_{\Psi_\alpha} \le M$ for some $\alpha \in (0,1] \cup \{2\}$ and let A be a symmetric d-tensor with vanishing diagonal such that $\|A\|_{\mathrm{HS}} \le 1$. Then*

$$\mathbb{E}\exp\Big(\frac{C_{d,\alpha}}{M^\alpha}|f_{d,A}(X)|^{\frac{\alpha}{d}}\Big) \le 2.$$

As in Theorem 5.2, the conclusion is equivalent to a $\Psi_{\alpha/d}$-norm estimate.

5.1 General results

In comparison to the aforementioned results, the main concentration inequalities provide more refined tail estimates. Here, we focus on the case $\alpha = 2/q$ for some $q \in \mathbb{N}$, which is sufficient for many applications, like products or powers of sub-Gaussian or sub-exponential random variables. The general case $\alpha \in (0,1]$ will be treated in Section 5.6. The results will be stated using a family of tensor-product matrix norms $\|A\|_{\mathcal{J}}$ for a d-tensor A and a partition $\mathcal{J} \in P_{qd}$ of $[qd]$. For the exact definitions, we refer to (5.11). Using these norms, we may formulate our first result for chaos-type functions.

Theorem 5.4. *Let $X_1, \dots, X_n$ be independent random variables such that for some $q \in \mathbb{N}$ and $M > 0$ we have $\|X_i\|_{\psi_{2/q}} \le M$, and let A be a symmetric d-tensor with vanishing diagonal. Consider $f_{d,A}(X)$ as in* (5.3)*. Then, for any $t \ge 0$,*

$$\mathbb{P}(|f_{d,A}(X)| \ge t) \le 2\exp\Big(-\frac{1}{C_{d,q}}\min_{\mathcal{J}\in P_{qd}}\Big(\frac{t}{M^d\|A\|_{\mathcal{J}}}\Big)^{\frac{2}{|\mathcal{J}|}}\Big).$$

To give an elementary example, consider the case $d = 1$ and $q = 2$. Here, $A = a = (a_1, \dots, a_n)$ is a vector, and $f_{1,A}(X) = \sum_{i=1}^n a_i(X_i - \mathbb{E}\,X_i)$ is a linear functional of random variables with sub-exponential tails, i. e. $\|X_i\|_{\psi_1} \le M$. It easily follows from the definition that $\|A\|_{\{1,2\}} = |a|$ (i. e. the Euclidean norm of a) and $\|A\|_{\{\{1\},\{2\}\}} = \max_i |a_i|$. As a consequence, for any $t \ge 0$

$$\mathbb{P}\Big(\big|\sum_{i=1}^n a_i(X_i - \mathbb{E}\,X_i)\big| \ge t\Big) \le 2\exp\Big(-\frac{1}{C}\min\Big(\frac{t^2}{M^2|a|^2}, \frac{t}{M\max_i |a_i|}\Big)\Big).$$

Hence, up to constants, we get back a classical result for the tails of a linear form in random variables with sub-exponential tails. For more general functions f and similar results under a Poincaré-type inequality, we refer to [BL97] (the first order case) and [GS19] (the higher order case).

Moreover, Theorem 5.4 can be used to give Hanson–Wright-type bounds for quadratic forms in $(2/q)$-sub-exponential random variables. Here we provide a sharpened version of Proposition 5.1.

Proposition 5.5. *Let $q \in \mathbb{N}$, $A = (a_{ij})$ be a symmetric $n \times n$ matrix and let $X_1, \ldots, X_n$ be independent, centered random variables with $\|X_i\|_{\Psi_{2/q}} \leq M$ and $\mathbb{E}\, X_i^2 = \sigma_i^2$. For any $t \geq 0$*

$$\mathbb{P}\Big(\big|\sum_{i,j} a_{ij} X_i X_j - \sum_{i=1}^n \sigma_i^2 a_{ii}\big| \geq t\Big) \leq 2 \exp\Big(-\frac{1}{C}\eta(A,q,t/M^2)\Big),$$

where

$$\eta(A,q,t) := \min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}, \Big(\frac{t}{\max_{i\in[n]}\|(a_{ij})_j\|_2}\Big)^{\frac{2}{q+1}}, \Big(\frac{t}{\|A\|_\infty}\Big)^{\frac{1}{q}}\Big).$$

Consequently, for any $x > 0$ we have with probability at least $1 - 2\exp(-x/C)$

$$|\langle X, AX\rangle - \mathbb{E}\langle X, AX\rangle| \leq M^2 \max\big(\sqrt{x}\|A\|_{\mathrm{HS}}, x\|A\|_{\mathrm{op}}, x^{\frac{q+1}{2}} \max_{i\in[n]}\|(a_{ij})_j\|_2, x^q\|A\|_\infty\big).$$

It is possible to replace $2/q$ by a general $\alpha \in (0,1] \cup \{2\}$ (see Section 5.6). In this case, we have to replace $2/(q+1)$ by $2\alpha/(2+\alpha)$ and $1/q$ by $\alpha/2$.

Remark. In comparison to the Hanson–Wright inequality (1.8) and Proposition 5.1, the more refined version contains two additional terms. The norms $\max_{i\in[n]}\|(a_{ij})_j\|_2$ and $\|A\|_\infty$ can no longer be written in terms of the eigenvalues of A (in contrast to $\|A\|_{\mathrm{HS}}$ and $\|A\|_{\mathrm{op}}$). Indeed, as we see later, we have $\max_{i\in[n]}\|(a_{ij})_j\|_2 = \|A\|_{2\to\infty}$, and $\|A\|_\infty = \max_{i,j}|\langle e_i, Ae_j\rangle|$ for the standard basis $(e_i)_i$ of $\mathbb{R}^n$. Moreover, the norms might have a very different scaling in n. For example, if $e = (1, \ldots, 1)$ and $A = ee^T - \mathrm{Id}$, then $\|A\|_{\mathrm{HS}} \sim \|A\|_{\mathrm{op}} \sim n$, $\max_{i\in[n]}\|(a_{ij})_j\|_2 \sim n^{1/2}$ and $\|A\|_\infty = 1$.

Remark. For the various forms of the Hanson–Wright inequality for quadratic forms in sub-Gaussian random variables we refer to [Ada15; ALM18; CY18; HW71; HKZ12; RV13; VW15; Wri73].

Finally, let us state the result for general polynomials in $(2/q)$-sub-exponential random variables. To fix some notation, if $f\colon \mathbb{R}^n \to \mathbb{R}$ is a function in $\mathcal{C}^D(\mathbb{R}^n)$, for $d \leq D$ we denote by $f^{(d)}$ the (symmetric) d-tensor of its d-th order partial derivatives.

Theorem 5.6. *Let $X_1, \ldots, X_n$ be independent random variables such that for some $q \in \mathbb{N}$ and $M > 0$ we have $\|X_i\|_{\psi_{2/q}} \leq M$. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree $D \in \mathbb{N}$. Then, for any $t \geq 0$, it holds*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\Big(-\frac{1}{C_{D,q}} \min_{d\in[D]} \min_{\mathcal{J}\in P_{qd}} \Big(\frac{t}{M^d\|\mathbb{E}f^{(d)}(X)\|_{\mathcal{J}}}\Big)^{\frac{2}{|\mathcal{J}|}}\Big).$$

Note that if $f(X) = f_{D,A}(X)$ as in (5.3), only the D-th order tensor gives a contribution, i. e. we retrieve Theorem 5.4. A variant of Theorem 5.6 for polynomials in independent random variables with $\|X_i\|_{\psi_\alpha} \le 1$ for any $\alpha \in (0,1]$ is derived in Section 5.6.

Remark. With the help of these inequalities, it is possible to prove many results on concentration of linear and quadratic forms in independent random variables scattered throughout the literature. For example, [NSU19, Lemma A.7] is an immediate consequence of Theorem 5.4 (combined with Lemma B.1 for $f(X,X') = \sum_{i=1}^n a_i X_i X_i'$). In a similar way, one can deduce [Yan+19, Lemma C.4] by applying Theorem 5.4 to the random variable $Z_i := X_i Y_i$, whenever (X_i, Y_i) is a vector with sub-exponential marginal distributions. More generally, one can consider a linear form (or higher order polynomial chaoses) in a product of k random variables $X_1, \ldots, X_k$ with sub-exponential tails, for which Lemma B.1 provides estimates for the $\Psi_{\frac{1}{k}}$ norm.

Lastly, the results in [EYY12, Appendix B] can be sharpened for $\alpha \in (0,1] \cup \{2\}$ by a more general version of Proposition 5.5, using the same arguments as in [RV13, Section 3] to treat complex-valued matrices. We omit the details.

5.2 Applications

In the following, we provide some applications of our main results. In particular, all the results in this section follow from either Proposition 5.1 or 5.5. For any collection of random variables $X_1, \ldots, X_n$ we write $X = (X_1, \ldots, X_n)$.

5.2.1 Euclidean norm of a vector with independent components

As a start, Proposition 5.1 can be used to prove concentration properties of the Euclidean norm of a linear transformation of a random vector X consisting of independent, normalized random variables with α-sub-exponential tails. We give two different forms thereof. The first form is inspired by the results in [RV13] for sub-Gaussian random variables.

Proposition 5.7. *Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbb{E} X_i = 0, \mathbb{E} X_i^2 = 1, \|X_i\|_{\Psi_\alpha} \le M$ for some $\alpha \in (0,1] \cup \{2\}$ and let $B \neq 0$ be an $m \times n$ matrix. For any $c > 0$ and any $t \ge c\|B\|_{\mathrm{HS}}$ we have*

$$\mathbb{P}\Big(|\|BX\|_2 - \|B\|_{\mathrm{HS}}| \ge t \Big) \le 2 \exp\Big(- \frac{\min(c^{2-\alpha},1)}{CM^4 \|B\|_{\mathrm{op}}^\alpha} t^\alpha \Big). \tag{5.4}$$

Note that in the case $\alpha = 2$ the constant c is not present on the right hand side of (5.4) and thus we can choose any $t \ge 0$, which is exactly the statement of [RV13, Theorem 2.1]. In the general case, we need to restrict t to the order $\|B\|_{\mathrm{HS}}$.

Proof. First off, it suffices to prove the inequality for a matrix B such that $\|B\|_{\mathrm{HS}} = 1$ and $t \ge c$, since the general case follows by considering $\widetilde{B} := B\|B\|_{\mathrm{HS}}^{-1}$.

Let us apply Proposition 5.1 to the matrix $A := B^T B$. An easy calculation shows that $\operatorname{tr}(A) = \operatorname{trace}(B^T B) = \|B\|_{\mathrm{HS}}^2 = 1$, so that we have

$$\begin{aligned}
\mathbb{P}\Big(|\|BX\|_2^2 - 1| \geq t\Big) &\leq 2\exp\Big(-\frac{1}{CM^4}\min\Big(\frac{t^2}{\|B\|_{\mathrm{op}}^2}, \Big(\frac{t}{\|B\|_{\mathrm{op}}^2}\Big)^{\frac{\alpha}{2}}\Big)\Big) \\
&= 2\exp\Big(-\frac{1}{CM^4\|B\|_{\mathrm{op}}^{\alpha}}\min\Big(\frac{t^{2-\alpha}}{\|B\|_{\mathrm{op}}^{2-\alpha}}t^{\alpha}, t^{\frac{\alpha}{2}}\Big)\Big) \\
&\leq 2\exp\Big(-\frac{\min(c^{2-\alpha}t^{\alpha}, t^{\frac{\alpha}{2}})}{CM^4\|B\|_{\mathrm{op}}^{\alpha}}\Big) \\
&\leq 2\exp\Big(-\frac{\min(c^{2-\alpha},1)}{CM^4\|B\|_{\mathrm{op}}^{\alpha}}\min(t^{\alpha}, t^{\frac{\alpha}{2}})\Big).
\end{aligned} \tag{5.5}$$

Here, in the first step we have used the estimates $\|A\|_{\mathrm{HS}}^2 \leq \|B\|_{\mathrm{op}}^2\|B\|_{\mathrm{HS}}^2 = \|B\|_{\mathrm{op}}^2$ and $\|A\|_{\mathrm{op}} \leq \|B\|_{\mathrm{op}}^2$ as well as the fact that by Lemma B.2, $\mathbb{E}X_i^2 = 1$ for any i implies $M \geq C_\alpha > 0$. The second inequality follows from $t \geq c \geq c\|B\|_{\mathrm{op}}$ and the third inequality is a consequence of $\min(c^{2-\alpha}t^{\alpha}, t^{\frac{\alpha}{2}}) \geq \min(c^{2-\alpha},1)\min(t^{\alpha}, t^{\frac{\alpha}{2}})$.

Now, as in [RV13], we use the inequality $|z-1| \leq \min(|z^2-1|, |z^2-1|^{1/2})$, giving for any $t \geq 0$

$$\mathbb{P}\Big(|\|BX\|_2 - 1| \geq t\Big) \leq \mathbb{P}\Big(|\|BX\|_2^2 - 1| \geq \max(t,t^2)\Big). \tag{5.6}$$

Hence, a combination of (5.5), (5.6) and $\min(\max(r,r^2), \max(r^{1/2},r)) = r$ yields for $t > c$

$$\mathbb{P}\Big(|\|BX\|_2 - 1| \geq t\Big) \leq 2\exp\Big(-\frac{\min(c^{2-\alpha},1)}{CM^4\|B\|_{\mathrm{op}}^{\alpha}}t^{\alpha}\Big).$$

□

The next corollary provides an alternative estimate for $\|BX\|_2$ if the X_i are sub-exponential.

Corollary 5.8. *Let $X_1, \ldots, X_n$ be independent, centered random variables satisfying $\|X_i\|_{\Psi_1} \leq M$ and $\mathbb{E}X_i^2 = \sigma_i^2$. For an $n \times n$ matrix B with real entries let $A = B^T B = (a_{ij})$. Then, for any $x > 0$, with probability at least $1 - 2\exp(-x/C)$ we have*

$$\|BX\|_2^2 \leq \sum_{i=1}^n \sigma_i^2 \sum_{j=1}^n b_{ji}^2 + M^2 \max\big(\sqrt{x}\|A\|_{\mathrm{HS}}, x\|A\|_{\mathrm{op}}, x^{3/2}\max_{i\in[n]}\|(a_{ij})_j\|_2, x^2\|A\|_\infty\big).$$

Corollary 5.8 can be compared to various bounds on the norms of $\|BX\|_2$ for a sub-Gaussian vector X (see [HKZ12] or [Ada15]). For example, if the sub-Gaussian constant is 1, in the sub-Gaussian case, with probability at least $1 - \exp(-x)$ it holds

$$\|BX\|_2^2 \leq \operatorname{tr}(B^T B) + 2\|B^T B\|_{\mathrm{HS}}\sqrt{x} + 2\|B^T B\|_{\mathrm{op}}x.$$

In Corollary 5.8 we have similar terms corresponding to $\sqrt{x}$ and x, whereas in the sub-exponential case we need two additional terms to account for the heavier tails of its components.

Proof. Define the quadratic form

$$Z := \|BX\|_2^2 = \langle BX, BX\rangle = \langle X, B^T BX\rangle = \langle X, AX\rangle.$$

Using Proposition 5.5 with the matrix A gives with probability $1 - 2\exp(-x/C)$

$$|Z - \mathbb{E}\, Z| \le \max\Big(\sqrt{x}\|A\|_{\mathrm{HS}}, x\|A\|_{\mathrm{op}}, x^{3/2} \max_{i=1,\dots,n} \|A_{i\cdot}\|_2, x^2\|A\|_\infty\Big).$$

From these inequalities and $|x| \ge x$ the claim easily follows by taking the square root. Note that $\mathbb{E}\, Z = \mathbb{E}\langle X, AX\rangle = \sum_{i=1}^n \sigma_i^2 \sum_{j=1}^n b_{ji}^2$. □

5.2.2 Projections and distance to a fixed subspace

It is possible to apply Proposition 5.5 to any matrix A associated to an orthogonal projection onto some lower-dimensional subspace. In these cases, the norms can be explicitly calculated and do not depend on the structure of the subspace, but merely on its dimension. This leads to the following application, where we replace a fixed projection by a random one.

Corollary 5.9. *Let $X_1, \dots, X_n$ be independent random variables satisfying $\mathbb{E}\, X_i = 0, \mathbb{E}\, X_i^2 = \sigma_i^2$ and $\|X_i\|_{\Psi_1} \le M$. Furthermore, let $m < n$ and P be the (random) orthogonal projection onto an m-dimensional subspace of $\mathbb{R}^n$, distributed according to the Haar measure on the Grassmanian manifold $G_{m,n}$, and independent of X. For any $x > 0$, with probability at least $1 - 2\exp(-x/C)$, we have*

$$\Big| \|PX\|_2^2 - \frac{m}{n}\sum_{i=1}^n \sigma_i^2 \Big| \le M^2 \max\Big(\sqrt{xm}, x^2\Big).$$

Proof. This is an application of Proposition 5.5. To see

$$\mathbb{E}\|PX\|_2^2 = \frac{m}{n}\sum_{i=1}^n \sigma_i^2,$$

we use [Ver18, Lemma 5.3.2] conditionally on X, i. e. we have

$$\mathbb{E}\|PX\|_2^2 = \mathbb{E}\,\mathbb{E}\left(\|PX\|_2^2 \mid X\right) = \frac{m}{n}\,\mathbb{E}\|X\|_2^2 = \frac{m}{n}\sum_{i=1}^n \sigma_i^2.$$

Moreover, for any projection P onto an m-dimensional subspace, one can see that $\|P\|_{\mathrm{HS}}^2 = \sum_{i=1}^n \lambda_i(P)^2 = m$. Moreover, we clearly have

$$\|P\|_\infty \le \max_{i=1,\dots,n} \|(p_{ij})_j\|_2 \le \|P\|_{2\to 2} = 1,$$

which finishes the proof. □

A very similar result which follows from Proposition 5.7 is the following variant of [RV13, Corollary 3.1]. We use the notation $d(X,E) = \inf_{e\in E} d(X,e)$ for the distance between an element X and a subset E of a metric space (M,d).

Corollary 5.10. *Let $X_1,\dots,X_n$ be independent random variables satisfying $\mathbb{E}\, X_i = 0, \mathbb{E}\, X_i^2 = 1$ and $\|X_i\|_{\Psi_\alpha} \le M$ for some $\alpha \in (0,1] \cup \{2\}$, and let E be a subspace of $\mathbb{R}^n$ of dimension d. For any $t \ge \sqrt{n-d}$ we have*

$$\mathbb{P}\Big(|d(X,E) - \sqrt{n-d}| \ge t\Big) \le 2\exp\Big(-\frac{t^\alpha}{CM^4}\Big).$$

Proof. This follows from Proposition 5.7 exactly as in [RV13, Corollary 3.1]. □

5.2.3 Spectral bound for a product of a fixed and a random matrix

We can also extend the second application in [RV13] to any set of α-sub-exponential random variables. For some $m \times n$ matrix B define the stable rank as $r(B) \coloneqq \|B\|^2_{mathrmHS}/\|B\|_{\mathrm{op}}$.

Proposition 5.11. *Let B be a fixed $m \times N$ matrix and let $G = (g_{ij})$ be a $N \times n$ random matrix with independent entries satisfying $\mathbb{E}\, g_{ij} = 0, \mathbb{E}\, g_{ij}^2 = 1$ and $\|g_{ij}\|_{\Psi_\alpha} \le M$ for some $\alpha \in (0,1]$. For any $u, v \ge 1$ with probability at least $1 - 2\exp(-u^\alpha r(B)^\alpha - v^\alpha n)$ we have*

$$\|BG\|_{\mathrm{op}} \le 4C_\alpha M^{4/\alpha}\Big(u\|B\|_{\mathrm{HS}} + vn^{1/\alpha}\|B\|_{\mathrm{op}}\Big).$$

Proof. We mimic the proof of [RV13, Theorem 3.2]. For any fixed $x \in S^{n-1}$ consider the linear operator $T : \mathbb{R}^{Nn} \to \mathbb{R}^m$ given by $T(G) = BGx$, and (by abuse of notation) write T for the matrix corresponding to this linear map in the standard basis. Using Proposition 5.7 applied to the matrix T we have

$$\mathbb{P}\Big(|\|BGx\|_2 - \|T\|_{\mathrm{HS}}| \ge t\Big) \le 2\exp\Big(-\frac{t^\alpha}{CM^4\|T\|^\alpha_{\mathrm{op}}}\Big).$$

Now, since $\|T\|_{\mathrm{HS}} = \|B\|_{\mathrm{HS}}$ and $\|T\|_{\mathrm{op}} \le \|B\|_{\mathrm{op}}$, this yields for any $t \ge \|B\|_{\mathrm{HS}}$

$$\mathbb{P}\Big(\|BGx\|_2 > \|B\|_{\mathrm{HS}} + t\Big) \le 2\exp\Big(-\frac{t^\alpha}{CM^4\|B\|^\alpha_{\mathrm{op}}}\Big).$$

If we define $t = (2CM^4)^{1/\alpha}\Big(u\|B\|_{\mathrm{HS}} + (\log(5)+1)^{1/\alpha} vn^{1/\alpha}\|B\|_{\mathrm{op}}\Big)$ for arbitrary $u,v \ge 1$ and use the inequality $2(r+s)^\alpha \ge r^\alpha + s^\alpha$ valid for all $r, s \ge 0$, we obtain

$$\begin{aligned}\mathbb{P}\Big(\|BGx\|_2 > \|B\|_{\mathrm{HS}} + t\Big) &\le 2\exp\Big(-\frac{u^\alpha\|B\|^\alpha_{\mathrm{HS}} + v^\alpha n(\log(5)+1)\|B\|^\alpha_{\mathrm{op}}}{\|B\|^\alpha_{\mathrm{op}}}\Big)\\ &\le 2\exp\Big(-u^\alpha r(B)^\alpha - v^\alpha n - v^\alpha n\log(5)\Big)\end{aligned}$$

$$\leq 5^{-n} 2 \exp\Big(- u^\alpha r(B)^\alpha - v^\alpha n\Big).$$

The last step is a covering argument as in [RV13]. Choose a 1/2-covering $\mathcal{N}$ (satisfying $|\mathcal{N}| \leq 5^n$, see [Ver12, Lemma 5.2]) of the unit sphere in $\mathbb{R}^n$, and note that a union bound gives

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{x \in \mathcal{N}} \|BGx\|_2 \leq \|B\|_{\mathrm{HS}} + t\Big) &\geq 1 - \sum_{x \in \mathcal{N}} \mathbb{P}\Big(\|BGx\|_2 > \|B\|_{\mathrm{HS}} + t\Big) \\ &\geq 1 - 2\exp\Big(- u^\alpha r(B)^\alpha - v^\alpha n\Big).\end{aligned}$$

Now, [Ver12, Lemma 5.3] yields

$$\|BG\|_{\mathrm{op}} \leq 2 \max_{x \in \mathcal{N}} \|BGx\|_2 \leq 2(\|B\|_{\mathrm{HS}} + t),$$

from which the assertion easily follows by upper bounding and simplifying the expression $2\|B\|_{\mathrm{HS}} + 2t$. □

5.2.4 Concentration properties for fixed design linear regression

It is possible to extend the example of the fixed design linear regression in [HKZ12] to the situation of a sub-exponential noise (instead of sub-Gaussian).

To this end, let $y_1, \ldots, y_n \in \mathbb{R}^d$ be fixed vectors (commonly called *design vectors*), $Y = (y_1, \ldots, y_n)$ (the $d \times n$ *design matrix*) and assume that the $d \times d$ matrix $\Sigma = n^{-1} \sum_{i=1}^n y_i y_i^T$ is invertible; in this case, define $B := n^{-1} \Sigma^{-1/2} Y \in M(d \times n)$. Let $X_1, \ldots, X_n$ be independent random variables with $\|X_i\|_{\Psi_1} \leq M$ and define

$$\beta := n^{-1} \sum_{i=1}^n \mathbb{E}\, X_i \Sigma^{-1} y_i \quad \text{and} \quad \hat{\beta}(X) := n^{-1} \sum_{i=1}^n X_i \Sigma^{-1} y_i.$$

β is the coefficient vector of the least expected squared error and $\hat{\beta}(X)$ is its ordinary least squares estimator (given the observation X). The quality of the estimator $\hat{\beta}$ can be judged by the excess loss

$$R(X) = \|\Sigma^{1/2}(\hat{\beta}(X) - \beta)\|_2^2 = \sum_{i,j} a_{ij}(X_i - \mathbb{E}\, X_i)(X_j - \mathbb{E}\, X_j),$$

where $A = (a_{ij}) = B^T B = n^{-2} Y^T \Sigma^{-1} Y$, as can be shown by elementary calculations. Observe that this is a quadratic form in X_i with coefficients depending on the vectors y_i. Thus, Proposition 5.5 yields the following corollary.

Corollary 5.12. *In the above setting, for any $x > 0$ the inequality*

$$|R(X)| \leq 4M^2 \max\Big(\sqrt{x}\|A\|_{\mathrm{HS}}, x\|A\|_{\mathrm{op}}, x^{3/2} \max_{i=1,\ldots,n} \|(a_{ij})_j\|_2, x^2 \|A\|_\infty\Big)$$

holds with probability at least $1 - 2\exp(-x/C)$.

5.2.5 Special cases

It is possible to apply all results to random variables having a Poisson distribution, i. e. $X_i \sim \mathrm{Poi}(\lambda_i)$ for some $\lambda_i \in (0,\infty)$. By using the moment generating function of the Poisson distribution, it is easily seen that

$$\|X_i\|_{\Psi_1} = \frac{1}{\log\Big(\log(2)\lambda_i^{-1} + 1\Big)} =: g(\lambda_i).$$

The function g is increasing and satisfies $g(x) \sim \log(1/x)$ (for $x \to 0$) and $g(x) \sim x/\log(2)$ (for $x \to \infty$). More generally, if the random variable $|X|$ has a moment generating function $\varphi_{|X|}$ in a neighborhood of 0, it can be used to explicitly calculate the Ψ_1-norm. Indeed, we have $\mathbb{E}\exp(|X|/t) = \varphi_{|X|}(t^{-1})$, and so $\|X\|_{\Psi_1} = 1/\varphi_{|X|}^{-1}(2)$.

Thus, as a special case of Proposition 5.5, we obtain the following corollary.

Corollary 5.13. *Let $X_i \sim \mathrm{Poi}(\lambda_i)$, $B := g(\max_{i\in[n]} \lambda_i)$ and $A = (a_{ij})$ be a symmetric $n \times n$ matrix. We have for any $t \geq 0$*

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{i,j} a_{ij}X_iX_j - \sum_{i=1}^n a_{ii}\lambda_i\Big| \geq B^2 t\Big) \\
&\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}, \Big(\frac{t}{\max_i\|(a_{ij})_j\|_2}\Big)^{\frac{2}{3}}, \Big(\frac{t}{\|A\|_\infty}\Big)^{\frac{1}{2}}\Big)\Big) \\
&\leq 2\exp\Big(-\frac{1}{C}\min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \Big(\frac{t}{\|A\|_{\mathrm{op}}}\Big)^{\frac{1}{2}}\Big)\Big).
\end{aligned}$$

For Poisson chaos of arbitrary order $d \in \mathbb{N}$, one may derive similar results by evaluating Theorem 5.4 or Corollary 5.22 (both for $\alpha = 1$). Note though that already for $d = 1$, we lose a logarithmic factor in the exponent. However, we are not aware of any more refined fluctuation estimates for $d \geq 2$.

Another interesting example of a sub-exponential random variable arises in stochastic geometry. If $K \subseteq \mathbb{R}^n$ is an isotropic, convex body and X is distributed according to the cone measure on K, then $\|\langle X,\theta\rangle\|_{\Psi_1} \leq c$ for some constant c and any $\theta \in S^{n-1}$. For the details and the proof we refer to [PTT19, Lemma 5.1].

5.3 The multilinear case: Proof of Theorem 5.4

To begin with, let us introduce some notation. Let $\mathbf{i} = (i_1, \ldots, i_d) \in [n]^d$ be a multiindex. For any subset $C \subseteq [d]$ with cardinality $|C| > 1$, we may introduce the "generalized diagonal" of $[n]^d$ with respect to C by

$$\{\mathbf{i} \in [n]^d : i_k = i_l \text{ for all } k,l \in C\}. \tag{5.7}$$

This notion of generalized diagonals naturally extends to d-tensors $A = (a_{\mathbf{i}})_{\mathbf{i}\in[n]^d}$ (obviously, the generalized diagonal of A with respect to C is the set of coefficients

$a_{\mathbf{i}}$ such that $\mathbf{i}$ lies on the generalized diagonal of $[n]^d$ with respect to C). If $d = 2$ and $C = \{1,2\}$, this gives back the usual notion of the diagonal of an $n \times n$ matrix. Moreover, write

$$[n]^{\underline{d}} := \{\mathbf{i} \in [n]^d : i_1, \dots, i_d \text{ are pairwise different}\}.$$

If A, B are d-tensors, we define $\langle A, B\rangle = \sum_{\mathbf{i}\in[n]^d} a_{\mathbf{i}} b_{\mathbf{i}}$. Given a set of d vectors $v^1, \dots, v^d \in \mathbb{R}^n$, we write $v^1 \cdots v^d$ for the outer product

$$(v^1 \cdots v^d)_{i_1 \dots i_d} := \prod_{j=1}^d v^j_{i_j}.$$

In particular, we may regard A as a multilinear form by setting $A(v^1, \dots, v^d) := \langle A, v^1 \cdots v^d\rangle$ for any $v^1, \dots, v^d \in \mathbb{R}^n$. The latter idea may be generalized by noting that any partition $\mathcal{J} = \{J_1, \dots, J_k\}$ of $[d]$ induces a partition of the space of d-tensors as follows. Identify the space of all d-tensors with $\mathbb{R}^{n^d}$ and decompose

$$\mathbb{R}^{n^d} \cong \bigotimes_{i=1}^k \mathbb{R}^{n^{J_i}} \cong \bigotimes_{i=1}^k \bigotimes_{j\in J_i} \mathbb{R}^n. \tag{5.8}$$

For any $x = x^{(1)} \otimes \ldots \otimes x^{(k)}$, the identification with a d-tensor is given by $x_{\mathbf{i}} = \prod_{l=1}^k x^{(l)}_{\mathbf{i}_{J_l}}$. For example, for $d = 4$ and $\mathcal{I} = \{\{1,4\}, \{2,3\}\}$ we have two matrices x,y and $x_{J_1,J_2,J_3,J_4} = x_{J_1 J_4} y_{J_2 J_3}$. Using this representation, any d-tensor A can be trivially identified with a linear functional on $\mathbb{R}^{n^d}$ via the standard scalar product, i. e.

$$Ax = A\left(x^{(1)} \otimes \ldots \otimes x^{(k)}\right) = \langle A, x^{(1)} \otimes \ldots \otimes x^{(k)}\rangle = \sum_{\mathbf{i}\in[n]^d} a_{\mathbf{i}} \prod_{l=1}^k x^{(l)}_{\mathbf{i}_{J_l}}.$$

These identifications give rise to a family of tensor-product matrix norms: for any partition $\mathcal{J} \in P_d$, define a norm on the space (5.8) by

$$\|x\|_{\mathcal{J}} := \|x^{(1)} \otimes \ldots \otimes x^{(k)}\|_{\mathcal{J}} := \max_{i=1,\dots,k} \|x^{(i)}\|_2.$$

Now, we may define $\|A\|_{\mathcal{J}}$ as the the operator norm with respect to $\|\cdot\|_{\mathcal{J}}$:

$$\|A\|_{\mathcal{J}} = \sup_{\|x\|_{\mathcal{J}}\le 1} |Ax|. \tag{5.9}$$

This family of tensor norms agrees with the definitions in [Lat06] and [AW15] (among others).

Next we extend these definitions to a family of norms $\|A\|_{\mathcal{J}}$ where A is a d-tensor but $\mathcal{J} \in P_{qd}$ for some $q \in \mathbb{N}$. To this end, we first embed A into the space of qd-tensors as follows. We divide $\mathbf{i} \in [n]^{qd}$ into d consecutive blocks with q

indices in each block $(i_1, \dots, i_q), (i_{q+1}, \dots, i_{2q}), \dots$ and only consider such indices for which all elements of these blocks take the same value. In fact, this is an intersection of d "generalized diagonals". More formally, we let $e_q(A)$ the qd-tensor given by

$$(e_q(A))_{\mathbf{i}} := \begin{cases} a_{i_1 i_{q+1} i_{2q+1} \dots i_{(k-1)q+1}} & \text{if } i_{kq+j} = i_{kq+1} \ \forall k = 0, \dots, d-1 \ \forall j = 2, \dots, q \\ 0 & \text{else.} \end{cases} \tag{5.10}$$

Now we set

$$\|A\|_{\mathcal{J}} := \|e_q(A)\|_{\mathcal{J}}. \tag{5.11}$$

For $q = 1$, this definition trivially agrees with (5.9).

Example. As we have mentioned before, the case $d = 1$ and $q = 2$ is easy to visualize. Given a vector $a = (a_1, \dots, a_n)$ we have $e_2(a) = \operatorname{diag}(a_1, \dots, a_n)$, where diag denotes a diagonal matrix. More generally, any $q > 2$ gives rise to a q-tensor $A = (a_{ijk})$ with $a_{iii} = a_i$ for all $i \in [n]$ and 0 otherwise.

Remark 5.14. The norms (5.11) are monotone with respect to the underlying partition in the following sense. For any two partitions $\mathcal{I} = \{I_1, \dots, I_\mu\}$ and $\mathcal{J} = \{J_1, \dots, J_\nu\}$ of $[qd]$, we say that $\mathcal{I}$ is finer than $\mathcal{J}$ (and write $\mathcal{I} \preccurlyeq \mathcal{J}$) if for any $j \in [\mu]$ there is a $k \in [\nu]$ such that $I_j \subseteq J_k$. If $\mathcal{I} \preccurlyeq \mathcal{J}$, we have $\|A\|_{\mathcal{I}} \le \|A\|_{\mathcal{J}}$. In particular, we always have

$$\|A\|_{\{\{1\},\dots,\{qd\}\}} \le \|A\|_{\mathcal{J}} \le \|A\|_{\{1,\dots,qd\}}. \tag{5.12}$$

In view of (5.12), the two "extreme" norms corresponding to the coarsest and the finest partition of $[qd]$ deserve special attention. Firstly, it is elementary that

$$\|A\|_{\{1,\dots,qd\}} = \|e_q(A)\|_{\mathrm{HS}} = \|A\|_{\mathrm{HS}} = \Big(\sum_{\mathbf{i} \in [n]^d} a_{\mathbf{i}}^2 \Big)^{1/2}. \tag{5.13}$$

Secondly, we have by Lemma 5.17

$$\|A\|_{\{\{1\},\dots,\{qd\}\}} = \|e_q(A)\|_{\mathrm{op}} = \begin{cases} \|A\|_{\mathrm{op}} & q = 1 \\ \max_{i,j} |a_{ij}| & q \ge 2 \end{cases},$$

To prove Theorem 5.4, we furthermore need two auxiliary results. The first one compares the moments of sums of random variables with $(2/q)$-sub-exponential decay to moments of Gaussian polynomials and the second one provides the estimates for multilinear forms in Gaussian random variables.

Lemma 5.15 (Lemma 5.4 in [AW15])**.** *For any positive integer k and any $p \ge 2$, if $Y_1, \dots, Y_n$ are independent symmetric random variables with $\|Y_i\|_{\psi_{2/k}} \le M$, then*

$$\Big\| \sum_{i=1}^n a_i Y_i \Big\|_p \le C_k M \Big\| \sum_{i=1}^n a_i g_{i1} \cdots g_{ik} \Big\|_p,$$

where g_{ij} are independent standard normal random variables.

Theorem 5.16 (Theorem 1 in [Lat06])**.** *Let $A = (a_{\mathbf{i}})_{\mathbf{i} \in [n]^d}$ be a d-tensor, and let $G_1, \ldots, G_d$ be i. i. d. standard Gaussian random vectors in $\mathbb{R}^n$. Then, for every $p \geq 2$,*

$$C_d^{-1} \sum_{\mathcal{J} \in P_d} p^{|\mathcal{J}|/2} \|A\|_{\mathcal{J}} \leq \|\langle A, G_1 \cdots G_d \rangle\|_p \leq C_d \sum_{\mathcal{J} \in P_d} p^{|\mathcal{J}|/2} \|A\|_{\mathcal{J}}.$$

As usual, in the proof of Theorem 5.4 we show L^p-estimates for $f_{d,A}(X)$. Recall that the concentration inequalities follow easily by Proposition 2.10.

Proof of Theorem 5.4. For simplicity, we always write $f \coloneqq f_{d,A}(X)$. Moreover, without loss of generality, we may assume the X_i to be centered.

Let $X^{(1)}, \ldots, X^{(d)}$ be independent copies of the random vector X. Take a set of i. i. d. Rademacher variables $(\varepsilon_i^{(j)})_{i \in [n], j \in [d]}$, which are independent of the $(X^{(j)})_j$. By standard decoupling and symmetrization inequalities (see [PG99, Theorem 3.1.1] and [PG99, Lemma 1.2.6]),

$$\|f\|_p \leq C_d \Big\| \sum_{\mathbf{i} \in [n]^{\underline{d}}} a_{i_1,\ldots,i_d} X_{i_1}^{(1)} \cdots X_{i_d}^{(d)} \Big\|_p \leq C_d \Big\| \sum_{\mathbf{i} \in [n]^{\underline{d}}} a_{i_1,\ldots,i_d} \varepsilon_{i_1}^{(1)} X_{i_1}^{(1)} \cdots \varepsilon_{i_d}^{(d)} X_{i_d}^{(d)} \Big\|_p.$$

An iteration of Lemma 5.15 together with $\|X_i\|_{\psi_{2/q}} \leq M$ hence leads to

$$\|f\|_p \leq \; C_d M^d \Big\| \sum_{\mathbf{i} \in [n]^{\underline{d}}} a_{i_1,\ldots,i_d} (g_{i_1,1}^{(1)} \cdots g_{i_1,q}^{(1)}) \cdots (g_{i_d,1}^{(d)} \cdots g_{i_d,q}^{(d)}) \Big\|_p.$$

Here, $(g_{i,k}^{(j)})$ is an array of i. i. d. standard Gaussian random variables. Rewriting (recall (5.10)) and applying Theorem 5.16 yields

$$\|f\|_p \leq \; C_d M^d \|\langle e_q(A), \otimes_{j=1}^d \otimes_{k=1}^q (g_{i,k}^{(j)})_{i \leq n} \rangle\|_p \leq \; C_d M^d \sum_{\mathcal{J} \in P_{qd}} p^{|\mathcal{J}|/2} \|A\|_{\mathcal{J}}.$$

□

5.4 Hanson–Wright-type inequality: Proof of Proposition 5.5

The main task in the proof of Proposition 5.5 is explicitly calculating the norms.

Lemma 5.17. *For any d-tensor A and $q \geq 2$ we have*

$$\|A\|_{\{\{1\},\ldots,\{qd\}\}} = \|A\|_\infty = \max_{i_1,\ldots,i_d} |a_{i_1,\ldots,i_d}|.$$

Proof. Write $\mathcal{J} = \{\{1\}, \ldots, \{qd\}\}$. We have

$$\begin{aligned}
\|A\|_{\mathcal{J}} &= \sup\Big\{\Big| \sum_{i_1 \ldots i_{qd}} (e_q(A))_{i_1,\ldots,i_{qd}} x^1_{i_1} \cdots x^{qd}_{i_{qd}} \Big| : \|x^j\|_2 \leq 1 \text{ for all } j = 1, \ldots, qd\Big\} \\
&= \sup\Big\{\Big| \sum_{i_1,\ldots,i_d} a_{i_1,\ldots,i_d} x^1_{i_1} \cdots x^q_{i_1} x^{q+1}_{i_2} \cdots x^{2q}_{i_2} \cdots x^{(d-1)q+1}_{i_d} \cdots x^{qd}_{i_d} \Big| : \|x^j\|_2 \leq 1\Big\} \\
&\leq \|A\|_\infty \sup\Big\{ \sum_{i_1,\ldots,i_d} |x^1_{i_1} x^2_{i_1}| \cdots |x^{(d-1)q+1}_{i_d} x^{(d-1)q+2}_{i_d}| : \|x^j\|_2 \leq 1\Big\} \\
&\leq \|A\|_\infty.
\end{aligned}$$

In the third step, we have iteratively used that for x^j with $\|x^j\|_2 \leq 1$ we also have $|x^j_i| \leq 1$, and applied the Cauchy–Schwarz inequality d times in the last step.

To obtain the lower bound, let $l_1, \ldots, l_d$ be the index which achieves the maximum. Let $x^1 = \ldots = x^q = \delta_{l_1}$, $x^{q+1} = \ldots = x^{2q} = \delta_{l_2}$ and so on, so that

$$\|A\|_{\mathcal{J}} \geq |a_{l_1 \cdots l_d}| = \|A\|_\infty.$$

□

The following easy observation helps with calculating the norms $\|\cdot\|_{\mathcal{J}}$. For any partition $\mathcal{J} = \{J_1, \ldots, J_k\} \in P_{[qd]}$ we write $\widetilde{\mathcal{J}} = \{\widetilde{J}_1, \ldots, \widetilde{J}_k\}$ for

$$\widetilde{J}_j = \{i \in [n] : J_j \cap \{q(i-1)+1, \ldots, qi\} \neq \emptyset\}. \tag{5.14}$$

That is, the sets $\widetilde{J}_j$ indicate which of the d q-blocks intersect J_j. Note that $\cup_j \widetilde{J}_j = [d]$, but $\widetilde{\mathcal{J}}$ needs not be a partition of $[d]$. In fact, some sets I may even appear more than once (with a slight abuse of notation, we choose to keep the set notation in this case anyway). Note that Remark 5.14 extends from partitions to decompositions (all definitions remain valid, even in case of some sets appearing multiple times). Nevertheless, we have by definition

$$\|A\|_{\mathcal{J}} = \|A\|_{\widetilde{\mathcal{J}}} := \sup\Big\{ \sum_{i_1,\ldots,i_d} a_{i_1 \ldots i_d} \prod_{j=1}^k x^{(j)}_{\mathbf{i}_{\widetilde{J}_j}} : \|x^{(j)}_{\mathbf{i}_{\widetilde{J}_j}}\|_2 \leq 1\Big\}, \tag{5.15}$$

i. e. the norm does not depend on $\mathcal{J}$, but on its “projection” $\widetilde{\mathcal{J}}$. We use this observation in the next lemma to calculate the norms $\|A\|_{\mathcal{J}}$ for quadratic forms (i. e. $d = 2$) and any $q \geq 2$.

Lemma 5.18. *Let $A = (a_{ij})$ be a symmetric matrix, $q \geq 2$ and $\mathcal{J}$ be a partition of $[2q]$.*

(1) *If $\widetilde{\mathcal{J}}$ contains $\{1,2\}$ two or more times, then $\|A\|_{\mathcal{J}} = \|A\|_\infty$.*

(2) *If $\widetilde{\mathcal{J}}$ contains $\{1,2\}$ and $\{1\}$ and $\{2\}$, then $\|A\|_{\mathcal{J}} = \|A\|_\infty$.*

(3) *If $\widetilde{\mathcal{J}} = \{\{1,2\}, \{j\}, \ldots, \{j\}\}$ ($j = 1$ or $j = 2$), then $\|A\|_{\mathcal{J}} = \max_{i \in [n]} \|a_{i\cdot}\|_2$.*

(4) *If* $\widetilde{\mathcal{J}}$ *comprises* l *times* $\{1\}$ *and* k *times* $\{2\}$ *for* $k \geq 2, l \geq 2$, *then* $\|A\|_{\mathcal{J}} = \|A\|_\infty$. *On the other hand, if* $l = 1, k \geq 2$ *or* $k = 1, l \geq 2$ *we have* $\|A\|_{\mathcal{J}} = \max_{i \in [n]} \|A_{i\cdot}\|_2$.

(5) *If* $\widetilde{\mathcal{J}} = \{\{1\}, \{2\}\}$, *then* $\|A\|_{\mathcal{J}} = \|A\|_{\mathrm{op}}$.

(6) *We have* $\|A\|_{\{[qd]\}} = \|A\|_{\mathrm{HS}}$.

Proof. To see (1), write $\widetilde{\mathcal{J}} = \{\widetilde{J}_1, \ldots, \widetilde{J}_k\}$, use the triangle inequality and the fact that $\|x\|_\infty \leq \|x\|_{\mathrm{HS}}$ for any tensor x:

$$\|A\|_{\mathcal{J}} = \sup\Big\{\sum_{i,j} a_{ij} \prod_{k=1}^{l} x^{(k)}_{\mathrm{i}_{\widetilde{J}_k}}\Big\} \leq \|A\|_\infty \sup\Big\{\sum_{i,j} |x_{ij}||y_{ij}|\Big\} \leq \|A\|_\infty,$$

where the supremum is taken over all unit vectors $x^{(k)}$. The lower bound follows from (5.12) and Lemma 5.17.

(2) follows immediately from $\widetilde{\mathcal{J}} \preccurlyeq \{\{1,2\}, \{1,2\}\}$.

(3) follows from the triangle and Cauchy–Schwarz inequality:

$$\begin{aligned}\|A\|_{\mathcal{J}} &\leq \sup\Big\{\sum_i \Big|\prod_{k=1}^{l} y_i^k\Big|\Big|\sum_j a_{ij} x_{ij}\Big|\Big\} \leq \sup\Big\{\sum_i \Big|\prod_{k=1}^{l} y_i^k\Big| \|(a_{ij})_j\|_2 \|x_{i\cdot}\|_2\Big\} \\ &\leq \max_i \|(a_{ij})_j\|_2 \sup\Big\{\Big|\prod_{k=1}^{l} y_i^k\Big| \|x_{i\cdot}\|_2\Big\} \leq \max_i \|(a_{ij})_j\|_2.\end{aligned}$$

The lower bound is obtained by choosing $y^1, \ldots, y^l$ as Dirac deltas on the row for which $\max_{i \in [n]} \|A_{i\cdot}\|$ is attained.

To see (4), note that the case $k \geq 2, l \geq 2$ is very similar to the second part. If $l = 1, k \geq 2$ or $k = 1, l \geq 2$, similar arguments as in the third part give for any $x, y^1, \ldots, y^l$ with norm at most one

$$\Big|\sum_{i,j} a_{ij} x_i \prod_k y_j^k\Big| \leq \sum_j \Big|\prod_k y_j^k\Big|\Big|\sum_i a_{ij} x_i\Big| \leq \sum_j \Big|\prod_k y_j^k \|(a_{ij})_j\|_2\Big| \leq \max_i \|(a_{ij})_j\|_2.$$

The lower bound again follows by choosing suitable Dirac deltas.

(5) and (6) are obvious from the definitions. □

Actually, we have the equality

$$\max_{i \in [n]} \|(a_{ij})_j\|_2 = \|A\|_{2 \to \infty},$$

where $\|A\|_{p \to q} \coloneqq \sup\left\{\|Ax\|_q : \|x\|_p \leq 1\right\}$. For the proof, see [CTP19, Proposition 6.1]. Especially this yields $\max_{i \in [n]} \|(a_{ij})_j\|_2 \leq \|A\|_{\mathrm{op}}$.

We are now ready to prove Proposition 5.5. Throughout the rest of this section, for a matrix A let us denote by A^{od} its off-diagonal and by A^{d} the diagonal part.

Proof of Proposition 5.5. Lemma 5.18 shows that we only need to consider the four norms $\|A\|_{\mathrm{HS}}, \|A\|_{\mathrm{op}}, \max_{i\in[n]}\|(a_{ij})_j\|_2$ and $\|A\|_\infty$. It is easy to see that $\|A\|_{\mathrm{HS}} \geq \|A\|_{\mathrm{op}} \geq \max_i\|(a_{ij})_j\|_2 \geq \|A\|_\infty$. Thus, we need to determine which partitions give rise to which norms.

The only partition producing the Hilbert–Schmidt norm is $\mathcal{J}_1 = \{[qd]\}$, with $|\mathcal{J}_1| = 1$. The operator norm appears for the decomposition $\mathcal{J}_2 = \{\{1,\ldots,q\},\{q+1,\ldots,2q\}\}$ with $|\mathcal{J}_2| = 2$. Moreover, it is easy to see that all partitions $\mathcal{J}_3$ of $[2q]$ giving rise to $\max_{i\in[n]}\|(a_{ij})_j\|_2$ satisfy $|\mathcal{J}_3| \in \{2,\ldots,q+1\}$. Finally, for all $k = 2,\ldots,2q$ there are partitions $\mathcal{J}_4$ such that $\|A\|_{\mathcal{J}_4} = \|A\|_\infty$.

Hence for a diagonal-free matrix A we have by simply plugging in the norms calculated in Lemmas 5.17 and 5.18 into Theorem 5.4

$$\mathbb{P}\Big(\Big|\sum_{i,j} a_{ij}(X_iX_j - \mathbb{E}\,X_i\,\mathbb{E}\,X_j)\Big| \geq t\Big) \leq 2\exp\Big(-\frac{1}{C}\eta(A,q,t/M^2)\Big), \tag{5.16}$$

where

$$\begin{aligned}\eta(A,q,t) &= \min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}, \min_{l=2,\ldots,q+1}\Big(\frac{t}{\max_i\|(a_{ij})_j\|_2}\Big)^{\frac{2}{l}}, \min_{l=2,\ldots,2q}\Big(\frac{t}{\|A\|_\infty}\Big)^{\frac{2}{l}}\Big)\\ &= \min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}, \Big(\frac{t}{\max_i\|(a_{ij})_j\|_2}\Big)^{\frac{2}{q+1}}, \Big(\frac{t}{\|A\|_\infty}\Big)^{\frac{1}{q}}\Big).\end{aligned}$$

In the last two terms, we can choose the largest l since we can assume that $\frac{t}{\|A\|_{\mathcal{J}}} \geq 1$ for any partition $\mathcal{J}$, as the minimum is achieved in $\frac{t^2}{\|A\|_{\mathrm{HS}}^2}$ otherwise.

For matrices with non-vanishing diagonal, we divide the quadratic form into an off-diagonal and a purely diagonal part, i. e.

$$\sum_{i,j} a_{ij}X_iX_j = \sum_{i,j} a_{ij}^{\mathrm{od}}X_iX_j + \sum_{i=1}^n a_{ii}^{\mathrm{d}}X_i^2.$$

For brevity, let us define $P(t) := \mathbb{P}\Big(\big|\sum_{i,j} a_{ij}X_iX_j - \sum_{i=1}^n \sigma_i^2 a_{ii}\big| \geq t\Big)$. Use the above decomposition and the subadditivity to obtain

$$P(t) \leq \mathbb{P}\Big(|\sum_{i,j} a_{ij}^{\mathrm{od}}X_iX_j| \geq t/2\Big) + \mathbb{P}\Big(|\sum_{i=1}^n a_{ii}^{\mathrm{d}}(X_i^2 - \sigma_i^2)| \geq t/2\Big) =: p_1(t) + p_2(t).$$

Equation (5.16) can be used to upper bound $p_1(t)$ as

$$p_1(t) \leq 2\exp\Big(-\frac{1}{C_2}\eta(A^{od},q,t/M^2)\Big). \tag{5.17}$$

The diagonal term can be treated by applying Theorem 5.4 for $d = 1$, $q = 4$ and $a = (a_{ii})_{i\in[n]}$. Moreover, it is easy to see that we have $\|a\|_{\{1,2,3,4\}} = \sum_i (a_{ii}^{\mathrm{d}})^2$ (cf.

(5.13)) and $\|a\|_{\mathcal{J}} = \|A^{\mathrm{d}}\|_\infty$ for any other decomposition $\mathcal{J}$. Consequently,

$$\begin{aligned} p_2(t) &\leq 2\exp\Big(-\frac{1}{C_1}\min\Big(\frac{t^2}{\|A^{\mathrm{d}}\|_{\mathrm{HS}}^2}, \frac{t}{\|A^{\mathrm{d}}\|_\infty}, \Big(\frac{t}{\|A^{\mathrm{d}}\|_\infty}\Big)^{2/3}, \Big(\frac{t}{\|A^{\mathrm{d}}\|_\infty}\Big)^{1/2}\Big)\Big) \\ &= 2\exp\Big(-\frac{1}{C_1}\eta_{1,A^{\mathrm{d}}}(t)\Big). \end{aligned} \tag{5.18}$$

Thus, by combining (5.17) and (5.18) we have

$$P(t) \leq 4\exp\left(-C\min(\eta(A^{\mathrm{od}},q,t), \eta_{1,A^{\mathrm{d}}}(t))\right).$$

Now it remains to lower bound the minimum by grouping the terms according to the different powers of t. This gives

$$p(t) \leq 4\exp\Big(-\frac{1}{C}\widetilde{\eta}(A,q,t/M)\Big),$$

where

$$\begin{aligned} &\widetilde{\eta}(A,q,t) \\ &:= \min\Big(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\max(\|A^{\mathrm{od}}\|_{\mathrm{op}}, \|A^{\mathrm{d}}\|_\infty)}, \Big(\frac{t}{\max_{i\in[n]}\|(a_{ij})_j\|_2}\Big)^{\frac{2}{q+1}}, \Big(\frac{t}{\|A\|_\infty}\Big)^{\frac{1}{q}}\Big). \end{aligned}$$

Lastly, from the characterization $\|A\|_{\mathrm{op}} := \sup_{x\in S^{n-1}}|\langle x,Ax\rangle|$ it can be easily seen that the inequalities $\|A^{\mathrm{d}}\|_\infty \leq \|A\|_{\mathrm{op}}$ and $\|A^{\mathrm{od}}\|_{\mathrm{op}} \leq 2\|A\|_{\mathrm{op}}$ hold, and the constant 4 can be changed to 2 by adjusting the constant in the exponent. □

5.5 The polynomial case: Proof of Theorem 5.6

Let us now treat the case of general polynomials $f(X)$ of total degree $D \in \mathbb{N}$. Before we start, we need to discuss some more properties of the norms $\|A\|_{\mathcal{J}}$. To this end, for two d-tensors A,B we let $A \circ B := (a_{\mathbf{i}}b_{\mathbf{i}})_{\mathbf{i}\in[n]^d}$ be their Hadamard product. For a set $C \subseteq [n]^d$ we may define "indicator matrices" 1_C by setting $1_C = (a_{\mathbf{i}})_{\mathbf{i}}$ with $a_{\mathbf{i}} = 1$ if $\mathbf{i} \in C$ and $a_{\mathbf{i}} = 0$ otherwise. If $|\mathcal{J}| > 1$, we do not have

$$\|A \circ 1_C\|_{\mathcal{J}} \leq \|A\|_{\mathcal{J}} \tag{5.19}$$

in general. However, [AW15, Lemma 5.2] shows a number of situations in which such an inequality does hold.

Lemma 5.19. *Let $A = (a_{\mathbf{i}})_{\mathbf{i}\in[n]^d}$ be a d-tensor.*

1. *If $C = \{\mathbf{i}\colon i_{k_1} = j_1, \ldots, i_{k_l} = j_l\}$ for some $1 \leq k_1 < \ldots < k_l \leq d$ ("generalized row"), then* (5.19) *holds.*
2. *If $C = \{\mathbf{i}\colon i_k = i_l\ \forall k,l \in K\}$ for some $K \subset [d]$ ("generalized diagonal"), then* (5.19) *holds.*

3. If $C_1, C_2 \subset [n]^d$ are such that (5.19) *holds, then so is $C_1 \cap C_2$.*

There is another situation in which a version of (5.19) holds. For any partition $\mathcal{K} = \{K_1, \ldots, K_a\} \in P_d$ we define

$$L(\mathcal{K}) = \{\mathbf{i} \in [n]^d \colon i_k = i_l \Leftrightarrow \exists j \colon k, l \in K_j\}. \tag{5.20}$$

That is, $L(\mathcal{K})$ is the set of those indices for which the partition into level sets is equal to $\mathcal{K}$.

Lemma 5.20. *Let $\mathcal{J} \in P_{qd}$, $\mathcal{K} \in P_d$ and A be a d-tensor. Then,*

$$\|A \circ 1_{L(\mathcal{K})}\|_{\mathcal{J}} \le 2^{|\mathcal{K}|(|\mathcal{K}|-1)/2} \|A\|_{\mathcal{J}}.$$

Proof. This is a generalization of [AW15, Corollary 5.3] which corresponds to the case $q = 1$. First note that by definition,

$$\|A \circ 1_{L(\mathcal{K})}\|_{\mathcal{J}} = \|e_q(A \circ 1_{L(\mathcal{K})})\|_{\mathcal{J}} = \|e_q(A) \circ e_q(1_{L(\mathcal{K})})\|_{\mathcal{J}}.$$

Therefore, it suffices to prove that for any qd-tensor B,

$$\|B \circ e_q(1_{L(\mathcal{K})})\|_{\mathcal{J}} \le 2^{|\mathcal{K}|(|\mathcal{K}|-1)/2} \|B\|_{\mathcal{J}}.$$

To see this, observe that $e_q(1_{L(\mathcal{K})})$ is the indicator matrix of a set C which can be written as an intersection of $|\mathcal{K}|$ generalized diagonals (with the cardinality of the underlying sets of indices in (5.7) always being an integer multiple of q) and $|\mathcal{K}|(|\mathcal{K}| - 1)/2$ sets of the form $\{\mathbf{i} : i_{kq+1} \neq i_{lq+1}\}$ for $k < l$. Recall that

$$\|B \circ 1_{\{i_{kq+1} \neq i_{lq+1}\}}\|_{\mathcal{J}} = \|B - B \circ 1_{\{i_{kq+1} = i_{lq+1}\}}\|_{\mathcal{J}} \le 2\|B\|_{\mathcal{J}},$$

using Lemma 5.19 (2) in the last step. As a consequence, the claim follows by applying Lemma 5.19 (2) again and a generalization of Lemma 5.19 (3). $\square$

Finally, it remains to note that [AW15, Lemma 5.1] can be generalized as follows.

Lemma 5.21. *Let A be a d-tensor and $v_1, \ldots, v_d \in \mathbb{R}^n$. Then, for any partition $\mathcal{J} \in P_{qd}$, $\|A \circ \otimes_{i=1}^d v_i\|_{\mathcal{J}} \le \|A\|_{\mathcal{J}} \prod_{i=1}^d \|v_i\|_\infty$.*

Proof. Recall equations (5.14) and (5.15). We have

$$\begin{aligned}
\|A \circ \otimes_{i=1}^d v_i\|_{\mathcal{J}} &= \sup\Big\{ \sum_{i_1,\ldots,i_{qd}} (e_q(A))_{i_1 \ldots i_{qd}} (e_q(\otimes_{i=1}^d v_i))_{i_1 \ldots i_{qd}} \prod_{j=1}^k x^{(j)}_{\mathbf{i}_{J_j}} : \|x^{(j)}_{\mathbf{i}_{J_j}}\|_2 \le 1 \Big\} \\
&= \sup\Big\{ \sum_{i_1,\ldots,i_d} a_{i_1 \ldots i_d} v_1^{i_1} \cdots v_d^{i_d} \prod_{j=1}^k x^{(j)}_{\mathbf{i}_{\tilde{J}_j}} : \|x^{(j)}_{\mathbf{i}_{\tilde{J}_j}}\|_2 \le 1 \Big\} \\
&\le \sup\Big\{ \sum_{i_1,\ldots,i_d} a_{i_1 \ldots i_d} \prod_{j=1}^k x^{(j)}_{\mathbf{i}_{\tilde{J}_i}} : \|x^{(j)}_{\mathbf{i}_{\tilde{J}_j}}\|_2 \le 1 \Big\} \prod_{i=1}^d \|v_i\|_\infty
\end{aligned}$$

$$= \sup\Big\{ \sum_{i_1,\dots,i_{qd}} (e_q(A))_{i_1\dots i_{qd}} \prod_{j=1}^{k} x^{(j)}_{\mathbf{i}_{J_j}} : \|x^{(j)}_{\mathbf{i}_{J_j}}\|_2 \le 1 \Big\} \prod_{i=1}^{d} \|v_i\|_\infty$$
$$= \|A\|_{\mathcal{J}} \prod_{i=1}^{d} \|v_i\|_\infty.$$

To see the third step, for each v_l we choose a set $\mathcal{J}_j$ such that $l \in \mathcal{J}_j$ and then define vectors $\widetilde{x}^{(j)}_{\mathbf{i}_{\widetilde{J}_j}}$ by multiplying $x^{(j)}_{\mathbf{i}_{\widetilde{J}_j}}$ by the components of the vectors v_l which were attributed to $\mathcal{J}_j$. In particular, this leads to $\|\widetilde{x}^{(j)}_{\mathbf{i}_{\widetilde{J}_j}}\|_2 \le \prod_l \|v_l\|_\infty \|x^{(j)}_{\mathbf{i}_{\widetilde{J}_j}}\|_2$, where the product is taken over all the vectors v_l which were attributed to $x^{(j)}_{\mathbf{i}_{\widetilde{J}_j}}$. □

Before we begin with the proof of the concentration results for general polynomials, let us give some definitions. Boldfaced letters always represent a vector (mostly a multiindex with integer components), and for any vector $\mathbf{i}$ let $|\mathbf{i}| \coloneqq \sum_j i_j$. For the sake of brevity we define

$$I_{m,d} \coloneqq \{(i_1,\dots,i_m) \in \mathbb{N}^m : |\mathbf{i}| = d\},$$
$$I_{m,\le d} \coloneqq \{(i_1,\dots,i_m) \in \mathbb{N}^m : |\mathbf{i}| \le d\}.$$

Given two vectors $\mathbf{i},\mathbf{k}$ of equal size, we write $\mathbf{k} \le \mathbf{l}$ if $k_j \le l_j$ for all j, and $\mathbf{k} < \mathbf{l}$ if $\mathbf{k} \le \mathbf{l}$ and there is at least one index such that $k_j < l_j$. Lastly, by $f \lesssim g$ we mean an inequality of the form $f \le C_{D,q} g$.

Proof of Theorem 5.6. We assume $M = 1$. For the general case, given random variables $X_1,\dots,X_n$ with $\|X_i\|_{\Psi_{2/q}} \le M$, define $Y_i \coloneqq M^{-1}X_i$. The polynomial $f = f(X)$ can be written as a polynomial $\widetilde{f} = \widetilde{f}(Y)$ by appropriately modifying the coefficients, i. e. multiplying each monomial by M^r, where r is its total degree. Now it remains to see that $\partial_{i_1\dots i_j}\widetilde{f}(Y) = M^j \partial_{i_1\dots i_j} f(X)$.

Step 1. First, we reduce the problem to generalizations of chaos-type functions (5.3). Indeed, by sorting according to the total grade, f may be represented as

$$f(x) = \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k}\in I_{\nu,d}} \sum_{\mathbf{i}\in [n]^{\underline{\nu}}} c^{(d)}_{(i_1,k_1),\dots,(i_\nu,k_\nu)} x_{i_1}^{k_1} x_{i_2}^{k_2} \cdots x_{i_\nu}^{k_\nu} + c_0,$$

where the constants satisfy $c^{(d)}_{(i_1,k_1),\dots,(i_\nu,k_\nu)} = c^{(d)}_{(i_{\pi_1},k_{\pi_1}),\dots,(i_{\pi_\nu},k_{\pi_\nu})}$ for any permutation $\pi \in \mathcal{S}_\nu$. As in [AW15], by rearranging and making use of the independence of $X_1,\dots,X_n$, this leads to the estimate

$$|f(X) - \mathbb{E}f(X)| \le \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k}\in I_{\nu,d}} \Big| \sum_{\mathbf{i}\in [n]^{\underline{\nu}}} a_{\mathbf{i}}^{\mathbf{k}} (X_{i_1}^{k_1} - \mathbb{E}X_{i_1}^{k_1}) \cdots (X_{i_\nu}^{k_\nu} - \mathbb{E}X_{i_\nu}^{k_\nu}) \Big|,$$

where

$$a_{\mathbf{i}}^{\mathbf{k}} = \sum_{m=\nu}^{D} \sum_{\substack{k_{\nu+1},\dots,k_m>0 \\ k_1+\dots+k_m\le D}} \sum_{\substack{i_{\nu+1},\dots,i_m \\ (i_1,\dots,i_m)\in[n]^{\underline{m}}}} \binom{m}{\nu} c^{(k_1+\dots+k_m)}_{(i_1,k_1),\dots,(i_m,k_m)} \prod_{\alpha=1}^{m} \mathbb{E} X_{i_\alpha}^{k_{i_\alpha}}.$$

Step 2. Note that $\|X_i^k\|_{\psi_{2/(qk)}} = \|X_i\|^k_{\psi_{2/q}} \le 1$. Thus, slightly modifying the proof of Theorem 5.4 (in particular, also using Lemma 5.15 for the non-linear terms), we obtain the estimate

$$\|f(X) - \mathbb{E}f(X)\|_p \lesssim \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k}\in I_{\nu,d}} \Big\| \sum_{\mathbf{i}\in[n]^{\underline{\nu}}} a_{\mathbf{i}}^{\mathbf{k}} (g^{(1)}_{i_1,1}\cdots g^{(1)}_{i_1,qk_1})\cdots(g^{(\nu)}_{i_\nu,1}\cdots g^{(\nu)}_{i_\nu,qk_\nu}) \Big\|_p.$$

Here, $(g^{(j)}_{i,k})$ is an array of i. i. d. standard Gaussian random variables.

Moreover, the family $(a_{\mathbf{i}}^{\mathbf{k}})_{\nu\in[d],k\in I_{\nu,d},i\in[n]^{\underline{\nu}}}$ gives rise to a d-tensor A_d as follows. Given any index $\mathbf{i} = (i_1,\dots,i_d)$ there is a unique number $r \in [d]$ of distinct elements $j_1,\dots,j_r$ with each j_l appearing exactly k_l times in $\mathbf{i}$. Consequently, we set $a_{i_1\dots i_d} := a^{(l_1,\dots,l_r)}_{j_1,\dots,j_r}$, and $A_d = (a_{\mathbf{i}})_{\mathbf{i}\in[n]^d}$. Note that this is well-defined due to the symmetry assumption.

For any $\mathbf{k} \in I_{\nu,d}$ denote by $\mathcal{K}(\mathbf{k}) = \mathcal{K}(k_1,\dots,k_\nu) \in P_d$ the partition which is defined by splitting the set $[d]$ into consecutive intervals of length $k_1,\dots,k_\nu$. In other words, $\mathcal{K}(\mathbf{k}) = \{K_1,\dots,K_\nu\}$ with $K_l = \{\sum_{i=1}^{l-1} k_i + 1, \sum_{i=1}^{l-1} k_i + 2, \dots, \sum_{i=1}^{l} k_i\}$, $l = 1,\dots,\nu$. Now, recalling the definitions of e_q (5.10) and of $L(\mathcal{K})$ (5.20), by rewriting and applying Lemma 5.19 we obtain

$$\begin{aligned}
\|f(X) - \mathbb{E}f(X)\|_p &\lesssim \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k}\in I_{\nu,d}} \|\langle e_q(A_d \circ 1_{L(\mathcal{K}(\mathbf{k}))}), \otimes_{j=1}^{\nu} \otimes_{k=1}^{qk_j} (g^{(j)}_{i,k})_{i\le n}\rangle\|_p \\
&\lesssim \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k}\in I_{\nu,d}} \sum_{\mathcal{J}\in P_{qd}} p^{|\mathcal{J}|/2} \|A_d \circ 1_{L(\mathcal{K}(k_1,\dots,k_\nu))}\|_{\mathcal{J}} \\
&\lesssim \sum_{d=1}^{D} \sum_{\mathcal{J}\in P_{qd}} p^{|\mathcal{J}|/2} \|A_d\|_{\mathcal{J}}.
\end{aligned} \tag{5.21}$$

Step 3. Next, we replace $\|A_d\|_{\mathcal{J}}$ by $\|\mathbb{E}f^{(d)}(X)\|_{\mathcal{J}}$. To this end, first note that for $\mathbf{i} \in [n]^d$ with distinct indices $j_1,\dots,j_\nu$ which are taken $l_1,\dots,l_\nu$ times, we have

$$\mathbb{E}\frac{\partial^d f}{\partial x_{i_1}\dots\partial x_{i_d}}(X) = \sum_{\mathbf{k}:\mathbf{k}\ge \mathbf{l}} \sum_{m=\nu}^{D} \sum_{\substack{k_{\nu+1},\dots,k_m>0 \\ k_1+\dots+k_m\le D}} \sum_{\substack{j_{\nu+1},\dots,j_m \\ (j_1,\dots,j_m)\in[n]^{\underline{m}}}}$$
$$\left(\binom{m}{\nu} \nu!\, c^{(k_1+\dots+k_m)}_{(j_1,k_1),\dots,(j_m,k_m)} \prod_{\alpha=1}^{\nu} \mathbb{E}X_{j_\alpha}^{k_\alpha - l_\alpha} \prod_{\alpha=\nu+1}^{m} \mathbb{E}X_{j_\alpha}^{k_\alpha} \prod_{\alpha=1}^{\nu} \frac{k_\alpha!}{(k_\alpha - l_\alpha)!}\right)$$

$$= \nu! l_1! \cdots l_\nu! a_{i_1,\dots,i_d} + R^{(d)}_{\mathbf{i}},$$

where the "remainder term" $R^{(d)}_{\mathbf{i}}$ corresponds to the set of indices $\mathbf{k}$ satisfying $\mathbf{k} > \mathbf{l}$. If $d = D$, we clearly have $R^{(d)}_{\mathbf{i}} = 0$, and therefore

$$\mathbb{E}\frac{\partial^D f}{\partial x_{i_1}\dots\partial x_{i_D}}(X) = \nu! l_1! \cdots l_\nu! a_{i_1\dots i_D} = \nu! |I_1|! \cdots |I_\nu|! a_{i_1\dots i_D}, \tag{5.22}$$

where $\mathcal{I} = \{I_1, \dots, I_\nu\}$ is the partition given by the level sets of the index $\mathbf{i}$. It follows that for any partition $\mathcal{J} \in P_{qD}$,

$$\|A_D\|_{\mathcal{J}} \le \sum_{\mathcal{K}\in P_D} \|A_D \circ 1_{L(\mathcal{K})}\|_{\mathcal{J}} \le \sum_{\mathcal{K}\in P_D} \|\mathbb{E}f^{(D)}(X) \circ 1_{L(\mathcal{K})}\|_{\mathcal{J}} \lesssim \|\mathbb{E}f^{(D)}(X)\|_{\mathcal{J}},$$

using the partition of unity $1 = \sum_{\mathcal{K}\in P_D} 1_{L(\mathcal{K})}$ and the triangle inequality in the first, equation (5.22) in the second and Lemma 5.20 in the last step.

The proof is now completed by induction. More precisely, in the next step we show that for any $d \in [D-1]$ and any partitions $\mathcal{I} = \{I_1, \dots, I_\mu\} \in P_d$, $\mathcal{J} = \{J_1, \dots, J_\nu\} \in P_{qd}$,

$$\|R^{(d)} \circ 1_{L(\mathcal{I})}\|_{\mathcal{J}} \lesssim \sum_{k=d+1}^{D} \sum_{\substack{\mathcal{K}\in P_{qk}\\ |\mathcal{K}|=|\mathcal{J}|}} \|A_k\|_{\mathcal{K}}. \tag{5.23}$$

Having (5.23) at hand, it follows by reverse induction and Lemma 5.20 that

$$\sum_{d=1}^{D} \sum_{\mathcal{J}\in P_{qd}} p^{|\mathcal{J}|/2} \|A_d\|_{\mathcal{J}} \lesssim \sum_{d=1}^{D} \sum_{\mathcal{J}\in P_{qd}} p^{|\mathcal{J}|/2} \|\mathbb{E}f^{(d)}(X)\|_{\mathcal{J}}.$$

Plugging this into (5.21) and applying Proposition 2.10 finishes the proof.

Step 4: To show (5.23), let us analyze the "remainder tensors" $R^{(d)}$ in more detail. To this end, fix $d \in [D-1]$ and partitions $\mathcal{I} = \{I_1, \dots, I_\nu\} \in P_d$, $\mathcal{J} = \{J_1, \dots, J_\mu\} \in P_{qd}$, and let $\mathbf{l}$ be the vector with $l_\alpha := |I_\alpha|$ (note that this implies $|\mathbf{l}| = d$). For any $\mathbf{k} \in I_{\nu,\le D}$ with $\mathbf{k} > \mathbf{l}$, we define a d-tensor $S^{(d,\mathbf{k})}_{\mathcal{I}} = (s^{(d,k_1,\dots,k_\nu)}_{\mathbf{i}})_{\mathbf{i}\in[n]^d} = (s^{(d)}_{\mathbf{i}})_{\mathbf{i}\in[n]^d}$ as follows:

$$s^{(d)}_{\mathbf{i}} = 1_{\mathbf{i}\in L(\mathcal{I})} \sum_{m=\nu}^{D} \sum_{\substack{k_{\nu+1},\dots,k_m>0\\ k_1+\dots+k_m\le D}} \sum_{\substack{j_{\nu+1},\dots,j_m\\ (j_1,\dots,j_m)\in[n]^{\underline{m}}}} \binom{m}{\nu} c^{(k_1+\dots+k_m)}_{(j_1,k_1),\dots,(j_m,k_m)} \prod_{\alpha=1}^{\nu} \mathbb{E}X_{j_\alpha}^{k_\alpha-l_\alpha} \prod_{\alpha=\nu+1}^{m} \mathbb{E}X_{j_\alpha}^{k_\alpha}$$

Here, we denote by j_α the value of $\mathbf{i}$ on the level set I_α. Clearly,

$$R^{(d)} \circ 1_{L(\mathcal{I})} = \sum_{\substack{\mathbf{k}\in I_{\nu,\le D}\\ \mathbf{k}>\mathbf{l}}} \nu! \frac{k_1}{(k_1-l_1)!} \cdots \frac{k_\nu}{(k_\nu-l_\nu)!} S^{(d,\mathbf{k})}_{\mathcal{I}}.$$

Therefore, it remains to prove that there is a partition $\mathcal{K} \in P_{q|\mathbf{k}|}$ with $|\mathcal{K}| = |\mathcal{J}|$ such that

$$\|S_{\mathcal{I}}^{(d,\mathbf{k})}\|_{\mathcal{J}} \lesssim \|A_{|\mathbf{k}|}\|_{\mathcal{K}}. \tag{5.24}$$

The tensor will be given by an appropriate embedding of the d-tensor $S_{\mathcal{I}}^{(d,\mathbf{k})}$. To this end, choose any partition $\widetilde{\mathcal{I}} = \{\widetilde{I}_1, \dots, \widetilde{I}_\nu\} \in P_{|\mathbf{k}|}$ with $|\widetilde{I}_\alpha| = k_\alpha$ and $I_\alpha \subset \widetilde{I}_\alpha$ for all α. Embedding the d-tensor $S_{\mathcal{I}}^{(d,\mathbf{k})}$ into the space of $|\mathbf{k}|$-tensors is done by defining a new tensor $\widetilde{S}^{|\mathbf{k}|} = (\widetilde{s}_{\mathbf{i}}^{|\mathbf{k}|})_{\mathbf{i}}$ given by

$$\widetilde{s}_{\mathbf{i}}^{|\mathbf{k}|} = s_{\mathbf{i}_{[d]}}^{(d)} 1_{\mathbf{i} \in L(\widetilde{\mathcal{I}})}.$$

We choose the partition $\mathcal{K} = \{K_1, \dots, K_\mu\}$ defined in the following way: for any j, we have $J_j \subset K_j$, so that it remains to assign the elements $r \in \{qd+1, \dots, q|\mathbf{k}|\}$ to the sets K_j. Write $r = \eta q + m$ for some $\eta \in \{d, \dots, |\mathbf{k}| - 1\}$ and $m \in [q]$. Since $\widetilde{\mathcal{I}}$ is a partition of $|\mathbf{k}|$, there is a unique $j \in [\nu]$ such that $\eta + 1 \in \widetilde{I}_j$. Take the smallest element t in $\widetilde{I}_j$ and add r to the same set as $\pi(r) \coloneqq (t-1)q + m$. Note that $I_j \subset \widetilde{I}_j$ implies $t \in [d]$, so that the procedure is well-defined.

We claim that

$$\|S_{\mathcal{I}}^{(d,|\mathbf{k}|)}\|_{\mathcal{J}} \le \|\widetilde{S}^{|\mathbf{k}|}\|_{\mathcal{K}}. \tag{5.25}$$

To see this, let $x^{(\beta)} = (x_{\mathbf{i}_{J_\beta}}^{(\beta)})_{\beta \in [\mu]}$ be a collection of vectors satisfying $\|x^{(\beta)}\|_2 \le 1$. This gives rise to another collection of unit vectors $y^{(\beta)} = (y_{\mathbf{i}_{K_\beta}}^{(\beta)})_{\beta \in [\mu]}$ defined by

$$y_{\mathbf{i}_{K_\beta}}^{(\beta)} = x_{\mathbf{i}_{K_\beta \cap [qd]}}^{(\beta)} \prod_{r \in K_\beta \setminus [qd]} 1_{i_r = i_{\pi(r)}}$$

Now, it follows that

$$\sum_{|\mathbf{i}_{[d]}| \le n} s_{\mathbf{i}_{[d]}}^{(d)} \prod_{\beta=1}^{\mu} x_{(e_q(\mathbf{i}))_{J_\beta}}^{(\beta)} = \sum_{|\mathbf{i}_{[|\mathbf{k}|]}| \le n} \widetilde{s}_{\mathbf{i}_{|\mathbf{k}|}}^{|\mathbf{k}|} \prod_{\beta=1}^{\mu} x_{(e_q(\mathbf{i}))_{J_\beta}}^{(\beta)} = \sum_{|\mathbf{i}_{[|\mathbf{k}|]}| \le n} \widetilde{s}_{\mathbf{i}_{[|\mathbf{k}|]}}^{(|\mathbf{k}|)} \prod_{\beta=1}^{\mu} y_{(e_q(\mathbf{i}))_{K_\beta}}^{(\beta)}.$$

These equations follow from the definition of the matrix $\widetilde{S}^{|\mathbf{k}|}$ and the fact that if $\mathbf{i} \in e_q(L(\widetilde{\mathcal{I}}))$, then for $r > qd$, $i_r = i_{\pi(r)}$, which implies $y_{\mathbf{i}_{K_\beta}}^{(\beta)} = x_{\mathbf{i}_{K_\beta \cap [qd]}}^{(\beta)} = x_{\mathbf{i}_{J_\beta}}^{(\beta)}$. As this holds true for any collection $x^{(\beta)}$, we obtain (5.25).

Finally, we prove

$$\|\widetilde{S}^{|\mathbf{k}|}\|_{\mathcal{K}} \lesssim \|A_{|\mathbf{k}|}\|_{\mathcal{K}} \tag{5.26}$$

for any partition $\mathcal{K} \in P_{q|\mathbf{k}|}$. To see this, note that if $\mathbf{i} \in L(\widetilde{\mathcal{I}})$, we have $\widetilde{s}_{\mathbf{i}}^{|\mathbf{k}|} = a_{\mathbf{i}}^{|\mathbf{k}|} \prod_{\alpha=1}^{\nu} \mathbb{E} X_{i_\alpha}^{k_\alpha - l_\alpha}$. As a consequence,

$$\widetilde{S}^{|\mathbf{k}|} = (A_{|\mathbf{k}|} \circ 1_{L(\widetilde{\mathcal{I}})}) \circ \otimes_{\alpha=1}^{|\mathbf{k}|} v_\alpha,$$

where the vectors v_α are defined by $v_\alpha = (\mathbb{E} X_i^{k_\alpha - l_\alpha})_{i \in [n]}$ if $\alpha \in \{\min I_1, \dots, \min I_\nu\}$

and $v_\alpha = (1, \ldots, 1)$, otherwise. In particular, we always have $\|v_\alpha\|_\infty \lesssim 1$, and therefore, by Lemma 5.21,

$$\|\widetilde{S}^{|\mathbf{k}|}\|_{\mathcal{K}} \lesssim \|A_{|\mathbf{k}|} \circ 1_{L(\widetilde{\mathcal{I}})}\|_{\mathcal{K}},$$

from where we easily arrive at (5.26) by applying Lemma 5.20.

Now, (5.24) follows by combining (5.25) and (5.26), which finishes the proof. □

5.6 The general sub-exponential case: $\alpha \in (0,1]$

Using slightly different techniques than in the proofs of Theorems 5.4 and 5.6, we may obtain concentration results for polynomials in independent random variables with bounded ψ_α-norms for any $\alpha \in (0,1]$. Here, the key difference is that we will not compare their moments to products of Gaussians but to Weibull variables.

To this end, we need some further notations. Let $A = (a_{\mathbf{i}})_{\mathbf{i} \in [n]^d}$ be a d-tensor and $I \subset [d]$ a set of indices. Then, for any $\mathbf{i}_I := (i_j)_{j \in I}$, we denote by $A_{\mathbf{i}_{I^c}} = (a_{\mathbf{i}})_{\mathbf{i}_{I^c}}$ the $(d - |I|)$-tensor defined by fixing i_j, $j \in I$. For instance, if $d = 4$, $I = \{1,3\}$ and $i_1 = 1, i_3 = 2$, then $A_{\mathbf{i}_{I^c}} = (a_{1j2k})_{jk}$.

For $I = [d]$, i. e. we fix all indices of $\mathbf{i}$, we interpret $A_{\mathbf{i}_{I^c}} = a_{\mathbf{i}}$ as the $\mathbf{i}$-th entry of A. Moreover, in this case, we assume that there is a single element $\mathcal{J} \in P(I^c)$ (which we may call the "empty" partition), and $\|A_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} = |a_{\mathbf{i}}|$ is just the Euclidean norm of $a_{\mathbf{i}}$. Finally, note that if $I = \emptyset$, $\mathbf{i}_I$ does not indicate any specification, and $A_{\mathbf{i}_{I^c}} = A$.

Using the characterization of the Ψ_α norms in terms of the growth of L^p norms (see Appendix B for details), [KL15, Corollary 2] yields a result similar to Theorem 5.4 for all $\alpha \in (0,1]$:

Corollary 5.22. *Let $X_1, \ldots, X_n$ be independent, centered random variables satisfying $\|X\|_{\Psi_\alpha} \le M$ for some $\alpha \in (0,1]$, A be a symmetric d-tensor with vanishing diagonal and consider $f_{d,A} = f_{d,A}(X)$ as in (5.3). We have for any $t \ge 0$*

$$\mathbb{P}\Big(|f_{d,A}| \ge t\Big) \le 2\exp\Big(-\frac{1}{C_{d,\alpha}} \min_{I \subset [d]} \min_{\mathcal{J} \in P(I^c)} \Big(\frac{t}{M^d \max_{\mathbf{i}_I} \|A_{\mathbf{i}_{I^c}}\|_{\mathcal{J}}}\Big)^{\frac{2\alpha}{2|I| + \alpha|\mathcal{J}|}}\Big).$$

The main goal of this section is to generalize Corollary 5.22 to arbitrary polynomials similarly to Theorem 5.6, which is the following result.

Theorem 5.23. *Let $X_1, \ldots, X_n$ be independent random variables with $\|X_i\|_{\psi_\alpha} \le M$ for some $\alpha \in (0,1]$ and $M > 0$. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a polynomial of total degree $D \in \mathbb{N}$. Then, for any $t \ge 0$, it holds*

$$\begin{aligned}&\mathbb{P}(|f(X) - \mathbb{E}f(X)| \ge t)\\ &\quad\le 2\exp\Big(-\frac{1}{C_{D,\alpha}} \min_{d \in [D]} \min_{I \subset [d]} \min_{\mathcal{J} \in P(I^c)} \Big(\frac{t}{M^d \max_{\mathbf{i}_I} \|(\mathbb{E}f^{(d)}(X))_{\mathbf{i}_{I^c}}\|_{\mathcal{J}}}\Big)^{\frac{2\alpha}{2|I| + \alpha|\mathcal{J}|}}\Big).\end{aligned}$$

To prove Theorem 5.23, note that one particular example of centered random variables with $\|X\|_{\Psi_\alpha} \le M$ is given by symmetric Weibull variables with shape parameter α (and scale parameter 1), i. e. symmetric random variables w with $\mathbb{P}(|w| \ge t) = \exp(-t^\alpha)$. In fact, [KL15, Example 3] especially implies the following analogue of Lemma 5.16. Here, we write $A \sim_{d,\alpha} B$ for a two-sided inequality $C_{d,\alpha}^{-1} A \le B \le C_{d,\alpha} A$.

Lemma 5.24. *Let $A = (a_{\mathbf{i}})_{\mathbf{i} \in [n]^d}$ be a d-tensor and $(w_i^j)_{i\in[n],j\in[d]}$, an array of i. i. d. Weibull variables with shape parameter $\alpha \in (0,1]$. Then, for every $p \ge 2$,*

$$\|\langle A, w^1 \otimes \ldots \otimes w^d \rangle\|_p \sim_{d,\alpha} \sum_{I \subset [d]} \sum_{\mathcal{J} \in P(I^c)} p^{|I|/\alpha + |\mathcal{J}|/2} \max_{\mathbf{i}_I} \|A_{\mathbf{i}_{I^c}}\|_{\mathcal{J}}.$$

Moreover, we need a replacement of Lemma 5.15. Here, we use Weibull instead of Gaussian random variables to compare the p-th moments:

Lemma 5.25. *For any $k \in \mathbb{N}$, any $\alpha > 0$ and any $p \ge 1$ the following holds. For any set of independent, symmetric random variables $Y_1, \ldots, Y_n$ satisfying $\|Y_i\|_{\psi_{\frac{\alpha}{k}}} \le M$ we have*

$$\Big\| \sum_{i=1}^n a_i Y_i \Big\|_p \le 2 c_{\alpha,k} M \Big\| \sum_{i=1}^n a_i w_{i_1} \cdots w_{ik} \Big\|_p,$$

where w_{ij} are symmetric i. i. d. Weibull random variables with shape parameter α and $c_{\alpha,k} := (k/(1-\log(2)))^{k/\alpha}$.

Proof. Due to homogeneity we assume $M = 1$, and for brevity we set $c := c_{\alpha,k}$. By Markov's inequality we have $\mathbb{P}(|Y_i| \ge t) \le 2\exp(-t^{\alpha/k})$ for any $i \in [n]$ and all $t \ge 0$.

The inclusion $\{c^{1/k}|w_{i1}| \ge t^{1/k}, \ldots, c^{1/k}|w_{ik}| \ge t^{1/k}\} \subseteq \{c|w_{i1} \cdots w_{ik}| \ge t\}$ holds for any $i \in [n]$ and $t \ge 0$. This yields for all $t \ge 1$

$$\begin{aligned}
\mathbb{P}(c|w_{i1} \cdots w_{ik}| \ge t) &\ge \prod_{j=1}^k \mathbb{P}(c^{1/k}|w_{ij}| \ge t^{1/k}) = \exp\Big(- k\Big(\frac{t}{c}\Big)^{\alpha/k}\Big) \\
&= \exp\Big(- (1-\log(2)) t^{\alpha/k}\Big) \ge 2\exp\Big(- t^{\alpha/k}\Big) \\
&\ge \mathbb{P}(|Y_i| \ge t),
\end{aligned}$$

where the second inequality requires the condition $t \ge 1$. Now the rest follows exactly as in [AW15, Proof of Lemma 5.4].

Alternatively, one can extend the inequality to all $t \ge 0$ by multiplying the left hand side by a constant. Indeed, it is easy to see (by observing $\mathbb{P}(c|w_{i_1} \cdots w_{ik}| \ge 1) \ge 2/e$) that for all $t \ge 0$ it holds

$$\mathbb{P}(|Y_i| \ge t) \le \frac{e}{2} \mathbb{P}(c|w_{i_1} \cdots w_{ik}| \ge t).$$

Thus, the contraction principle [Kwa87, Theorem 1] tells us that for any $p \geq 1$ we have

$$\Big\| \sum_{i=1}^n a_i Y_i \Big\|_p \leq \frac{e}{2} c_{\alpha,k} M \Big\| \sum_{i=1}^n a_i w_{i_1} \cdots w_{ik} \Big\|_p .$$

□

Our next goal is to adapt Lemmas 5.19, 5.20 and 5.21 to the "restricted" tensors $A_{\mathbf{i}_{I^c}}$. That is, we examine whether (a modification of) the inequality

$$\|(A \circ 1_C)_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \leq \|A_{\mathbf{i}_{I_c}}\|_{\mathcal{J}} \tag{5.27}$$

still holds in this situation, where $\mathcal{J}$ is a partition of I^c.

Lemma 5.26. *Let $A = (a_{\mathbf{i}})_{\mathbf{i} \in [n]^d}$ be a d-tensor, $I \subset [d]$ and $\mathbf{i}_I \in [n]^I$ fixed.*

1. *If $C = \{\mathbf{i} \colon i_{k_1} = j_1, \ldots, i_{k_l} = j_l\}$ for some $1 \leq k_1 < \ldots < k_l \leq d$ ("generalized row"), then* (5.27) *holds.*
2. *If $C = \{\mathbf{i} \colon i_k = i_l \ \forall k, l \in K\}$ for some $K \subset [d]$ ("generalized diagonal"), then* (5.27) *holds.*
3. *If $C_1, C_2 \subset [n]^d$ are such that* (5.27) *holds, then so is $C_1 \cap C_2$.*
4. *If $\mathcal{K} \in P_d$, then $\|(A \circ 1_{L(\mathcal{K})})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \leq 2^{|\mathcal{K}|(|\mathcal{K}|-1)/2} \|A_{\mathbf{i}_{I^c}}\|_{\mathcal{J}}$.*
5. *For any vectors $v_1, \ldots, v_d \in \mathbb{R}^n$, $\|(A \circ \otimes_{i=1}^d v_i)_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \leq \|A_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \prod_{i=1}^d \|v_i\|_\infty$.*

Proof. To see (1), we may assume that $\{k_1, \ldots, k_l\} \cap I = \emptyset$ (note that in the case $\{k_1, \ldots, k_l\} \cap I \neq \emptyset$, either the conditions are not compatible, in which case $(A \circ 1_C)_{i_{I^c}} = 0$, or we can remove some of the conditions and obtain a subset with $\{k_1, \ldots, k_{\tilde{l}}\} \cap I = \emptyset$). In this case, if C is a generalized row, then $(A \circ 1_C)_{\mathbf{i}_{I^c}} = A_{\mathbf{i}_{I^c}} \circ 1_{C'}$ for some generalized row C' in I^c. This proves (1).

If C is a generalized diagonal, we have to consider two situations. Assuming $K \cap I = \emptyset$, i. e. K is subset of I^c, we immediately obtain (2). On the other hand, if $K \cap I \neq \emptyset$, then $(A \circ 1_C)_{\mathbf{i}_{I^c}} = A_{\mathbf{i}_{I^c}} \circ 1_{C'}$ for some generalized *row* C' in I^c, readily leading to (2) again.

(3) is clear. To see (4), one may argue as in the proof of Lemma 5.20 (for $q = 1$), replacing Lemma 5.19 (2) and (3) by their analogues we just proved. Finally, an easy modification of the proof of Lemma 5.21 yields (5). □

We are now ready to prove Theorem 5.23. Here, we recall the notation used in the proof of Theorem 5.6, with the only difference that now, by $f \lesssim g$ we mean an inequality of the form $f \leq C_{D,\alpha} g$.

Proof of Theorem 5.23. We will follow the proof of Theorem 5.6. In particular, let us assume $M = 1$.

Step 1. Recall the inequality from the proof of Theorem 5.6

$$|f(X) - \mathbb{E}f(X)| \le \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k} \in I_{\nu,d}} \Big| \sum_{\mathbf{i} \in [n]^{\underline{\nu}}} a_{\mathbf{i}}^{\mathbf{k}} (X_{i_1}^{k_1} - \mathbb{E}X_{i_1}^{k_1}) \cdots (X_{i_\nu}^{k_\nu} - \mathbb{E}X_{i_\nu}^{k_\nu}) \Big|.$$

Step 2. Applying Lemma 5.26, we arrive at

$$\|f(X) - \mathbb{E}f(X)\|_p \lesssim \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k} \in I_{\nu,d}} \Big\| \sum_{\mathbf{i} \in [n]^{\underline{\nu}}} a_{\mathbf{i}}^{\mathbf{k}} (w_{i_1,1}^{(1)} \cdots w_{i_1,k_1}^{(1)}) \cdots (w_{i_\nu,1}^{(\nu)} \cdots w_{i_\nu,k_\nu}^{(\nu)}) \Big\|_p.$$

Here, $(w_{i,k}^{(j)})$ is an array of i. i. d. symmetric Weibull variables with shape parameter α. Now we may define d-tensors A_d as in the proof of Theorem 5.6. Similarly as in (5.21), rewriting and applying Lemma 5.24 and Lemma 5.26 (4) yields

$$\begin{aligned}
\|f(X) - \mathbb{E}f(X)\|_p &\lesssim \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k} \in I_{\nu,d}} \|\langle A_d \circ 1_{L(\mathcal{K}(k_1,\ldots,k_\nu))}, \otimes_{j=1}^{\nu} \otimes_{k=1}^{k_j} (w_{i,k}^{(j)})_{i \le n} \rangle\|_p \\
&\lesssim \sum_{d=1}^{D} \sum_{\nu=1}^{d} \sum_{\mathbf{k} \in I_{\nu,d}} \sum_{I \subset [d]} \sum_{\mathcal{J} \in P(I^c)} p^{\frac{|I|}{r} + \frac{|\mathcal{J}|}{2}} \max_{\mathbf{i}_I} \|(A_d \circ 1_{L(\mathcal{K}(\mathbf{k}))})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \\
&\lesssim \sum_{d=1}^{D} \sum_{I \subset [d]} \sum_{\mathcal{J} \in P(I^c)} p^{\frac{|I|}{r} + \frac{|\mathcal{J}|}{2}} \max_{\mathbf{i}_I} \|(A_d)_{\mathbf{i}_{I^c}}\|_{\mathcal{J}}.
\end{aligned}$$

Step 3. In the proof of Theorem 5.6 we have decomposed

$$\mathbb{E} \frac{\partial^d f}{\partial x_{i_1} \ldots \partial x_{i_d}}(X) = \nu! l_1! \cdots l_\nu! a_{i_1,\ldots,i_d} + R_{\mathbf{i}}^{(d)}$$

with a remainder tensor $R_{\mathbf{i}}^{(d)}$ corresponding to the set of indices $\mathbf{k}$ with $\mathbf{k} > \mathbf{l}$ and $R_{\mathbf{i}}^{(d)} = 0$ for $d = D$. Again, for any $I \subset [D]$ and any partition $\mathcal{J} \in P(I^c)$,

$$\begin{aligned}
\|(A_D)_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} &\le \sum_{\mathcal{K} \in P_D} \|(A_D \circ 1_{L(\mathcal{K})})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \le \sum_{\mathcal{K} \in P_D} \|(\mathbb{E}f^{(D)}(X) \circ 1_{L(\mathcal{K})})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \\
&\lesssim \|(\mathbb{E}f^{(D)}(X))_{\mathbf{i}_{I^c}}\|_{\mathcal{J}},
\end{aligned}$$

using Lemma 5.26 (4) in the last step. To complete the proof, we need to show that for any $d = 1, \ldots, D-1$, any $I \subset [d]$ and any partitions $\mathcal{I} \in P([d])$, $\mathcal{J} \in P([d] \backslash I)$,

$$\|(R^{(d)} \circ 1_{L(\mathcal{I})})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \lesssim \sum_{k=d+1}^{D} \sum_{\substack{\mathcal{K} \in P([k] \backslash I) \\ |\mathcal{K}| \ge |\mathcal{J}|}} \|(A_k)_{\mathbf{i}_{I^c}}\|_{\mathcal{K}}. \tag{5.28}$$

Actually, analyzing the proof one can see it is possible to restrict the second sum on the right-hand side to partitions $\mathcal{K}$ with $|\mathcal{K}| \in \{|\mathcal{J}|, |\mathcal{J}|+1\}$. Once having proven (5.28), it follows from reverse induction that

$$\sum p^{|I|/r+|\mathcal{J}|/2} \max_{\mathbf{i}_I} \|(A_d)_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \lesssim \sum p^{|I|/r+|\mathcal{J}|/2} \max_{\mathbf{i}_I} \|(\mathbb{E} f^{(d)}(X))_{\mathbf{i}_{I^c}}\|_{\mathcal{J}},$$

where $\sum = \sum_{d\in[D]} \sum_{I\subseteq[d]} \sum_{\mathcal{J}\in P(I^c)}$. Here, we use that for any $p \geq 2$ and any $|\mathcal{K}| \geq |\mathcal{J}|$ we have $p^{|\mathcal{J}|/2} \leq p^{|\mathcal{K}|/2}$. In view of Step 2 and Proposition 2.10, this finishes the proof.

Step 4. Recall the d-tensor $S_{\mathcal{I}}^{(d,\mathbf{k})} = (s_{\mathbf{i}}^{(d,k_1,\ldots,k_\nu)})_{\mathbf{i}\in[n]^d} = (s_{\mathbf{i}}^{(d)})_{\mathbf{i}\in[n]^d}$ and for any $\mathbf{k} \in I_{\nu,\leq D}$ with $\mathbf{k} > \mathbf{l}$ the $|\mathbf{k}|$-tensor $\widetilde{S}^{|\mathbf{k}|}$ from the proof of Theorem 5.6.

In the sequel, we fix the following items:

- $d \in [D-1]$ and $\mathbf{k}$ satisfying $|\mathbf{k}| > d$,
- a set $I \subset [d]$ and $\mathbf{i}_I \in [n]^I$,
- an admissible partition $\mathcal{I} \in P([d])$ and an associated extension $\widetilde{I} \in P([|\mathbf{k}|])$.

The notion of admissibility was not relevant in Theorem 5.6, as we have not fixed any indices I and values $\mathbf{i}_I \in [n]^I$. Here, it simply means that the level sets have to compatible with the fact that we have fixed some of the partial derivatives by I and $\mathbf{i}_I$. Also note that $\mathcal{I}$ is a partition of $[d]$ and not of $[d]\backslash I$, since it arises from level sets of partial derivatives and includes the those from I.

Our aim is to find a partition $\mathcal{K} \in P([|\mathbf{k}|]\backslash I)$ with $|\mathcal{K}| \in \{|\mathcal{J}|, |\mathcal{J}|+1\}$ such that

$$\|(S_{\mathcal{I}}^{(d,\mathbf{k})})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \lesssim \|(A_{|\mathbf{k}|})_{\mathbf{i}_{I^c}}\|_{\mathcal{K}}. \tag{5.29}$$

First, set $\mathcal{K} \coloneqq \mathcal{J} = \{J_1, \ldots, J_\mu\}$, so that it remains to assign the elements $r \in \{d+1, \ldots, |\mathbf{k}|\}$. This is done as follows: Select $r \in \{d+1, |\mathbf{k}|\}$, and do the following steps:

- choose $k \in \mathbb{N}$ such that $r \in \widetilde{I}_k$,
- take the smallest element $t =: \pi(r)$ in $\widetilde{I}_k$ (since $I_k \subset \widetilde{I}_k$, we have $t \in [d]$),
- if $t \in I^c$, there is a set $J_j \ni t$ and we add r to J_j; otherwise, we assign r to an "extra set" $K_{\mu+1}$

In particular, it may happen that $K_{\mu+1} = \emptyset$. In this case, we ignore $\beta = \mu+1$ in the rest of the proof.

Now we claim that it holds

$$\|(S_{\mathcal{I}}^{(d,|\mathbf{k}|)})_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \leq \|(\widetilde{S}^{|\mathbf{k}|})_{\mathbf{i}_{I^c}}\|_{\mathcal{K}}. \tag{5.30}$$

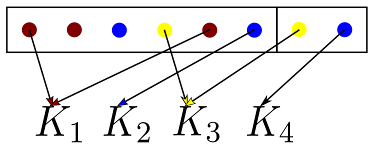

Figure 5.1: An illustration of the procedure of producing the partition $\mathcal{K}$. Here, we start with $d = 6$, $|\mathbf{k}| = 8$, $I = \{2,3\}$. The partition $\widetilde{\mathcal{I}}$ is indicated by colors, i. e. $\widetilde{\mathcal{I}} = \{\{1,2,5\}, \{3,6,8\}, \{4,7\}\}$. $\{8\}$ belongs to $K_4 = K_{\mu+1}$ since $\{3\} \in I$. Changing its color to yellow would produce a partition $\mathcal{K}$ with 3 subsets.

To see this, let $x = (x^{(\beta)})_{\beta \in [\mu]} = ((x^{(\beta)}_{\mathbf{i}_{J_\beta}}))_{\beta \in [\mu]}$ be a unit vector with respect to $\mathcal{J}$, i. e. $|x|_{\mathcal{J}} = \max_{\beta \in [\mu]} |x^{(\beta)}| \le 1$. We embed it in the unit ball with respect to $|x|_{\mathcal{K}}$ by defining $y = (y^{(\beta)})_{\beta \in [\mu+1]}$ via

$$y^{(\beta)}_{\mathbf{i}_{K_\beta}} = \begin{cases} x^{(\beta)}_{\mathbf{i}_{K_\beta \cap [d]}} \prod_{r \in K_\beta \setminus [d]} 1_{i_r = i_{\pi(r)}} & \beta \in [\mu] \\ \prod_{r \in K_{\mu+1}} 1_{i_r = i_{\pi(r)}} & \beta = \mu + 1. \end{cases}$$

As $y^{(\mu+1)}$ has a single non-zero entry, it is easy to see that $|y|_{\mathcal{K}} \le 1$. Moreover, by the definition of the matrix $\widetilde{S}^{|k|}$ and the fact that if $\mathbf{i} \in L(\widetilde{\mathcal{I}})$, then for $r > d$, $i_r = i_{\pi(r)}$, which implies $y^{(\beta)}_{\mathbf{i}_{K_\beta}} = x^{(\beta)}_{\mathbf{i}_{K_\beta \cap [d]}} = x^{(\beta)}_{\mathbf{i}_{J_\beta}}$ as well as $y^{(\mu+1)}_{\mathbf{i}_{K_{\mu+1}}} = 1$ we have

$$\langle (S^{(d,\mathbf{k})})_{\mathbf{i}_{I^c}}, \otimes_{\beta=1}^{\mu} x^{(\beta)} \rangle = \langle (\widetilde{S}^{(|k|)})_{\mathbf{i}_{I^c}}, \otimes_{\beta=1}^{\mu+1} y^{(\beta)} \rangle.$$

Hence, the supremum on the left hand side of (5.30) is taken over a subset of the unit ball with respect to $|x|_{\mathcal{K}}$. Finally, it remains to prove

$$\|(\widetilde{S}^{|\mathbf{k}|})_{\mathbf{i}_{I^c}}\|_{\mathcal{K}} \lesssim \|(A_{|\mathbf{k}|})_{\mathbf{i}_{I^c}}\|_{\mathcal{K}} \tag{5.31}$$

for any partition $\mathcal{K} \in P(I^c)$. This may be achieved as in the proof of Theorem 5.6, replacing Lemma 5.21 by Lemma 5.26 (5).

Combining (5.30) and (5.31) yields (5.29), which finishes the proof. □

Finally, we prove Proposition 5.1 and Theorem 5.2, from which Corollary 5.3 follows immediately.

Proof of Proposition 5.1. The case $\alpha \in (0,1]$ follows from the $d = 2$ case of Corollary 5.22. $\alpha = 2$ corresponds to the Hanson–Wright inequality. □

Proof of Theorem 5.2. Let $\alpha \in (0,1]$ and consider the bound given by Theorem 5.23. Fix any $d \in [D]$, and observe that for any $I \subset [d]$, any $\mathbf{i}_I$ and any $\mathcal{J} \in P(I^c)$ we have by (5.13)

$$\|(\mathbb{E} f^{(d)}(X))_{\mathbf{i}_{I^c}}\|_{\mathcal{J}} \le \|(\mathbb{E} f^{(d)}(X))_{\mathbf{i}_{I^c}}\|_{\mathrm{HS}} \le \|\mathbb{E} f^{(d)}(X)\|_{\mathrm{HS}}$$

as well as

$$\frac{\alpha}{d} \le \frac{2\alpha}{2|I| + \alpha|\mathcal{J}|} \le 2.$$

If $t \ge M^d \|\mathbb{E} f^{(d)}(X)\|_{\mathrm{HS}}$, this immediately yields the result. Otherwise, note that the tail bound given in Theorem 5.2 is trivial. (In fact, here one needs to ensure that $C_{D,\alpha}$ is sufficiently large, e. g. $C_{D,\alpha} \ge 1$.)

In a similar way, it is possible to derive the same results for $\alpha = 2/q$ and any $q \in \mathbb{N}$ from Theorem 5.6. From these results, the exponential moment bound follows by standard arguments, see for example [BGS19, Proof of Theorem 1.1]. $\square$

APPENDIX A

Approximate tensorization of entropy in finite spaces

The concentration results for finite spin systems were based on an approximate tensorization property of the entropy. Here, we reformulate and provide a complete proof of a result of Marton [Mar15] and moreover rewrite it in the terms of entropy of functions instead of relative entropy of measures. Let $\mathcal{X}$ be a finite set, $\mathcal{X}^n$ its n-fold product and fix a probability measure q^n on $\mathcal{X}^n$. Define the total variation distance, the relative entropy and the Wasserstein-2-type distance

$$\begin{aligned} d_{TV}(\mu,\nu) &\coloneqq \sup_{A \subseteq \mathcal{X}^n} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}^n} |\mu(\{x\}) - \nu(\{x\})|, \\ H(\mu \mid\mid \nu) &= \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu \quad \text{if} \quad \mu \ll \nu \\ W_2(\mu, \nu) &= \inf_{\pi \in C(\mu,\nu)} \Big(\sum_{i=1}^n \pi(x_i \neq y_i)^2 \Big)^{1/2}, \end{aligned}$$

where $C(\mu,\nu)$ is the set of all couplings of μ and ν, i. e. probability measures π on $\mathcal{X}^n \times \mathcal{X}^n$ with marginals μ and ν.

Remark. The infimum in the definition of W_2 is always attained, since $C(\mu,\nu)$ is a compact subset of $\mathcal{P}(\mathcal{X}^n \times \mathcal{X}^n)$ equipped with the weak topology and the map $\pi \mapsto \left(\sum_{i=1}^n \pi(x_i \neq y_i)^2\right)^{1/2}$ is lower semicontinuous. This fact and the gluing lemma for measures with a common marginal [AG13, Theorem 2.1] can be used to prove that W_2 is a distance function on $\mathcal{P}(\mathcal{X}^n)$, see for example [Vil09, Chapter 6] for a similar line of reasoning.

Denote by μ_i, ν_i the pushforward of μ and ν respectively of the projection onto the i-th coordinate. By the subadditivity of the square root (for the upper bound for W_2) as well as the fact that every $\pi \in C(\mu,\nu)$ induces a coupling of μ_i, ν_i via the projection onto the x_i and y_i coordinate (for the lower bound), we obtain

$$\sum_{i=1}^n d_{TV}^2(\mu_i, \nu_i) \le W_2^2(\mu,\nu) \le n d_{TV}(\mu,\nu). \tag{A.1}$$

We need the following lemma, which is a slight improvement of [Mar15, Lemma 2].

Lemma A.1. *Let q be a probability measure on a finite space $\mathcal{X}$ and $\beta_q := \inf_{x\in\mathcal{X}_+} q(x)$, where $\mathcal{X}_+ := \{x \in \mathcal{X} : q(x) > 0\}$. For any probability measure $p \ll q$ we have*

$$H(p \mid\mid q) \le 2\beta_q^{-1} d_{TV}^2(p,q).$$

Proof. The shifted logarithm $f(x) := \log(1+x)$ is a concave function on $(-1,\infty)$, so that for any $x \ge -1$ the inequality $f(x) \le f'(0)x = x$ holds. Rewrite $\frac{p}{q} = 1+\frac{p-q}{q}$ to obtain

$$\begin{aligned} H(p \mid\mid q) &= \sum_{x\in\mathcal{X}_+} q(x)\left(1+\frac{p(x)-q(x)}{q(x)}\right) f\left(\frac{p(x)-q(x)}{q(x)}\right) \\ &\le \sum_{x\in\mathcal{X}_+} q(x)\left(1+\frac{p(x)-q(x)}{q(x)}\right)\frac{p(x)-q(x)}{q(x)} = \sum_{x\in\mathcal{X}_+}\frac{(p(x)-q(x))^2}{q(x)} \\ &\le \beta_q^{-1}\sum_{x\in\mathcal{X}_+}(p(x)-q(x))^2 \le \beta_q^{-1} d_{TV}(p,q)\sum_{x\in\mathcal{X}}|p(x)-q(x)| \\ &= 2\beta_q^{-1} d_{TV}^2(p,q). \end{aligned}$$

□

Unfortunately, the factor $2\beta_q^{-1}$ cannot be removed. To see this, let $\mathcal{X} = \{0,1\}$ and $q(0) = 1-q(1) = \alpha$ for a fixed $\alpha \in (0,1/2)$. Now, considering the family of measures $p_\varepsilon(0) = \alpha+\varepsilon$, an easy calculation yields $d_{TV}^2(p_\varepsilon,q) = \varepsilon^2$ and $H(p_\varepsilon \mid\mid q) \sim 2\alpha^{-1}\varepsilon^2$.

A consequence of Lemma A.1 is that on finite spaces the relative entropy and the total variation distance are comparable with a constant depending on the (fixed) measure q, i. e. we have

$$d_{TV}^2(p,q) \le \frac{1}{2}H(p \mid\mid q) \le \beta_q^{-1} d_{TV}^2(p,q).$$

Here, the first inequality is the well-known Pinsker inequality (also known as Csiszár–Kullback–Pinsker inequality), see for example [Tsy09, Lemma 2.5].

Theorem A.2. *Let q^n be a probability measure on a finite space $\mathcal{X}^n$ with full support and set $\beta := \min_{i\in[n]}\min_{x\in\mathcal{X}^n} q_i(x_i \mid \overline{x}_i)$.*

(*i*) *Let p^n be a probability measure on $\mathcal{X}^n$. If for all subsets $I \subseteq [n]$ and all $\overline{y}_I$ we have*

$$W_2^2(p_I(\cdot \mid \overline{y}_I), q_I(\cdot, \overline{y}_I)) \le C\sum_{i\in I}\mathbb{E}_{p_I(\cdot|\overline{y}_I)}\, d_{TV}^2(p_i(\cdot \mid \overline{y}_i), q_i(\cdot \mid \overline{y}_i)), \qquad \text{(A.2)}$$

then the approximate tensorization property holds:

$$H(p^n \mid\mid q^n) \le \frac{C}{\beta}\sum_{i=1}^{n}\mathbb{E}_{\overline{p}_i}\, H(p_i(\cdot \mid \overline{y}_i) \mid\mid q_i(\cdot \mid \overline{y}_i)). \qquad \text{(A.3)}$$

In particular, if f denotes the density of p^n with respect to q^n, then this can be rewritten as

$$\operatorname{Ent}_{q^n}(f) \leq \frac{C}{\beta} \sum_{i=1}^{n} \int \operatorname{Ent}_{q_i(\cdot|\overline{y}_i)}(f(\overline{y}_i, \cdot)) dq^n(y). \tag{A.4}$$

(*ii*) *Assume that an interdependence matrix A of q^n satisfies $\|A\|_{2\to 2} < 1$. Then* (A.2) *holds with $C = (1 - \|A\|_{2\to 2})^{-2}$. In particular,* (A.3) *and* (A.4) *hold with the same constant.*

Proof. (i): We prove the theorem by induction. In the case $n = 1$ there is nothing to prove if one interprets $q_1(\cdot \mid \overline{y}_1) = q$. The disintegration theorem for the relative entropy (see for example [DZ10, Theorem D.13]) yields

$$H(p^n \mid\mid q^n) = \frac{1}{n} \sum_{i=1}^{n} H(p_i \mid\mid q_i) + \frac{1}{n} \sum_{i=1}^{n} \int H(\overline{p}_i(\cdot \mid y_i) \mid\mid \overline{q}_i(\cdot \mid y_i)) dp_i(y_i).$$

We bound the two terms separately. For the first term, using Lemma A.1, equations (A.1), (A.2) and Pinsker's inequality gives

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} H(p_i \mid\mid q_i) &\leq \frac{2}{\beta n} \sum_{i=1}^{n} d_{TV}^2(p_i, q_i) \leq \frac{2}{\beta n} W_2^2(p^n, q^n) \\
&\leq \frac{2C}{\beta n} \sum_{i=1}^{n} \mathbb{E}_{p^n} \, d_{TV}^2(p_i(\cdot \mid \overline{y}_i), q_i(\cdot \mid \overline{y}_i)) \\
&\leq \frac{C}{\beta n} \sum_{i=1}^{n} \mathbb{E}_{p^n} \, H(p_i(\cdot \mid \overline{y}_i) \mid\mid q_i(\cdot \mid \overline{y}_i)).
\end{aligned}$$

We apply the induction hypothesis to rewrite the second term. For each fixed $i \in [n]$ and $y_i \in \mathcal{X}$ we interpret $\widetilde{q} := \overline{q}_i(\cdot \mid y_i)$ as a measure on $\overline{\mathcal{X}}_i$ satisfying

$$\beta(\widetilde{q}) = \min_{j:j\neq i} \min_{x \in \overline{\mathcal{X}}_i} \frac{\overline{q}_i(x \mid y_i)}{\overline{q}_i(\overline{x}_j \mid y_i)} = \min_{j:j\neq i} \min_{x \in \overline{\mathcal{X}}_i} \frac{q^n(x, y_i)}{q^n(\overline{x}_j, y_i)} \geq \min_{j \in [n]} \min_{z \in \mathcal{X}^n} \frac{q^n(z)}{\overline{q_j}(\overline{z_j})} = \beta(q^n)$$

and (A.2) with the same constant C. To rewrite (A.3) let us denote by $y \in \overline{\mathcal{X}}_i$ a generic vector. A short calculation shows that the conditional probability of $\overline{p}_i(\cdot \mid y_i)$ with respect to the projection $\overline{pr}_j : \overline{\mathcal{X}}_i \to \overline{\mathcal{X}}_{ij}$ for some $j \neq i$ is given by $p_j(y_j \mid \overline{y}_j, y_i)$, which is the conditional probability of p^n given $(\overline{y}_j, y_i)$, and the same holds for $q(\cdot \mid y_i)$. Thus we obtain

$$H(p(\cdot \mid y_i) \mid\mid q(\cdot \mid y_i)) \leq \frac{C}{\beta} \sum_{j:j\neq i} \mathbb{E}_{p(\cdot|y_i)} \, H\big(p(\cdot \mid \overline{y}_j, y_i) \mid\mid q(\cdot \mid \overline{y}_j, y_i)\big).$$

Integration with respect to p_i and summation over i yields

$$\frac{1}{n}\sum_{i=1}^{n}\int H(\overline{p}_i(\cdot \mid y_i) \mid\mid \overline{q}_i(\cdot \mid y_i))dp_i(y_i) \leq \frac{C}{\beta}\frac{n-1}{n}\sum_{i=1}^{n}\mathbb{E}_{p^n} H(p_i(\cdot \mid \overline{y}_i) \mid\mid q_i(\cdot \mid \overline{y}_i)),$$

which combined with the first term proves the assertion.

(A.4) is a simple rewriting of (A.3), noting that as a consequence of the disintegration theorem (or in this case Bayes' theorem) we have

$$\frac{dp_i(\cdot \mid \overline{y}_i)}{dq_i(\cdot \mid \overline{y}_i)}(y_i) = \frac{f(\overline{y}_i, y_i)}{\int f(\overline{y}_i, x_i)dq_i(x_i \mid \overline{y}_i)}$$

and $\frac{d\overline{p}_i}{d\overline{q}_i}(\overline{x}_i) = \int f(\overline{x}_i, x_i)dq_i(x_i \mid \overline{x}_i)$.

(ii): See [Mar15, Theorem 2]. □

In [Mar15, Theorem 1] it is stated that using the quantity

$$\beta := \inf_{i=1,\ldots,n}\inf_{x\in\mathcal{X}^n : q^n(x)>0} q_i(x_i \mid \overline{x}_i)$$

one can deduce $q^n(pr_i(x) = x_i) \geq \beta$ for all x_i such that the LHS is nonzero. This is possible only if q^n has full support. A counterexample is given by the push-forward of a random uniform permutation under the map $\sigma \mapsto (\sigma_1, \ldots, \sigma_n)$, which satisfies $\beta = 1$.

Actually, we can modify the condition of Theorem A.2 to allow for measures without full support. This may be done by using the definition $\widetilde{\beta}$ as in (2.24) instead of β. As $\widetilde{\beta}$ is a uniform control, it can be used in any step of the induction, so that the theorem remains valid. Clearly the inequality (A.3) only holds for $p^n \ll q^n$ only, and (A.4) holds for positive functions vanishing outside the support of q^n.

APPENDIX B

Properties of Orlicz quasinorms

As mentioned in Chapter 5, the Orlicz norms defined in (5.2) satisfy the triangle inequality only for $\alpha \geq 1$. Indeed, the function $g_\alpha : x \mapsto \exp(x^\alpha)$ is not convex around 0 for $\alpha \in (0,1)$, see the following figure. A short calculation yields that g_α is convex in the region $[(\alpha^{-1} - 1)^{\alpha^{-1}}, \infty)$ only, and concave around 0.

However, for any $\alpha \in (0,1)$ this is still a quasinorm, which for many purposes is sufficient. We shall collect some elementary results on Orlicz quasinorms in this appendix. The first result is a Hölder-type inequality for the Ψ_α norms.

Lemma B.1. *Let $X_1, \ldots, X_k$ be random variables such that $\|X_i\|_{\Psi_{\alpha_i}} < \infty$ for some $\alpha_i \in (0,1]$ and let $t := (\sum_{i=1}^k \alpha_i^{-1})^{-1}$. Then $\|\prod_{i=1}^k X_i\|_{\Psi_t} < \infty$ and*

$$\Big\| \prod_{i=1}^{k} X_i \Big\|_{\Psi_t} \leq \prod_{j=1}^{k} \|X_i\|_{\Psi_{\alpha_i}}.$$

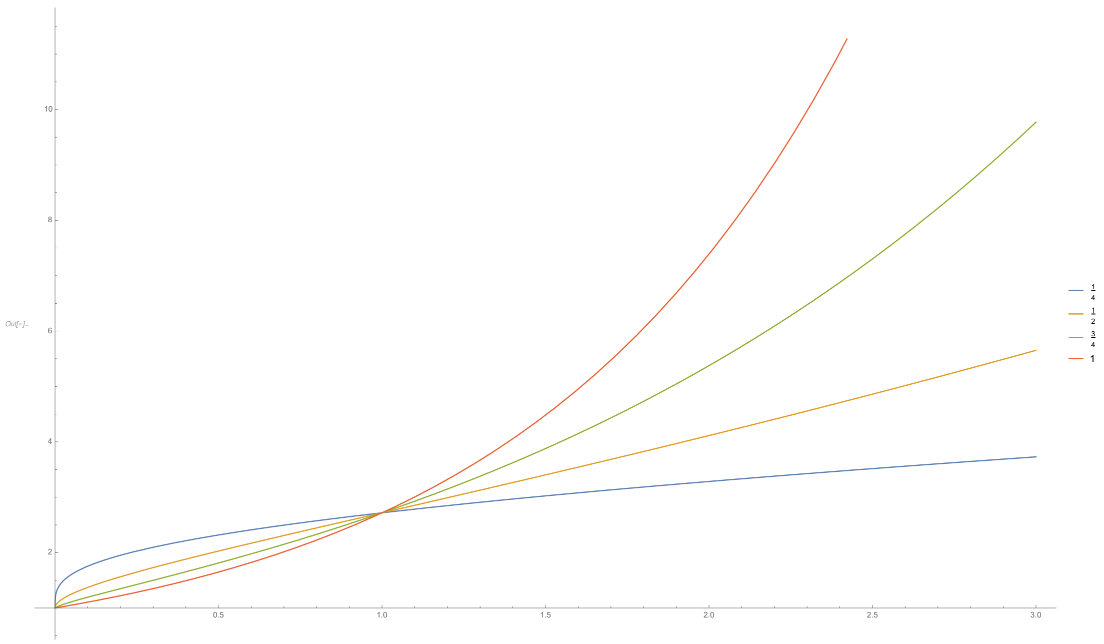

Figure B.1: The function $g_\alpha(x), x \in [0,3]$ for four different values of α.

Especially for $\alpha_i = \alpha$ *for all* $i \in [k]$ *this leads to* $\|\prod_{i=1}^k X_i\|_{\Psi_{\alpha/k}} \leq \prod_{i=1}^k \|X_i\|_{\Psi_\alpha}$.

The random variables $X_1, \ldots, X_k$ need not be independent, i. e. we can consider a random vector $X = (X_1, \ldots, X_k)$ with marginals having α-sub-exponential tails.

Proof. By homogeneity we can assume $\|X\|_{\Psi_{\alpha_i}} = 1$ for all $i \in [k]$. We need the general form of Young's inequality, i. e. for all $p_1, \ldots, p_k > 1$ satisfying $\sum_{i=1}^k p_i^{-1} = 1$ and any $x_1, \ldots, x_k \geq 0$ we have

$$\prod_{i=1}^k x_i \leq \sum_{i=1}^k p_i^{-1} x_i^{p_i},$$

which follows easily from the concavity of the logarithm. If we apply this to $p_i := \alpha_i t^{-1}$ and use the convexity of the exponential function, we obtain

$$\mathbb{E} \exp\Big(\prod_{i=1}^k |X_i|^t\Big) \leq \mathbb{E} \exp\Big(\sum_{j=1}^k p_i^{-1} |X_i|^{\alpha_i}\Big) \leq \sum_{j=1}^k p_i^{-1}\, \mathbb{E} \exp\Big(|X_i|^{\alpha_i}\Big) \leq 2.$$

This is however equivalent to $\|\prod_{i=1}^k X_i\|_{\Psi_t} \leq 1$. □

To state the other lemmas, for any $0 < \alpha < 1$ define

$$d_\alpha := (\alpha e)^{1/\alpha}/2 \qquad \text{and} \qquad D_\alpha := (2e)^{1/\alpha}.$$

Lemma B.2. *For any* $0 < \alpha < 1$ *we have*

$$d_\alpha \sup_{p \geq 1} \frac{\|X\|_p}{p^{1/\alpha}} \leq \|X\|_{\Psi_\alpha} \leq D_\alpha \sup_{p \geq 1} \frac{\|X\|_p}{p^{1/\alpha}}. \tag{B.1}$$

The statement of the lemma remains true for $\alpha \geq 1$, with (α-independent constants) $d_\alpha = 1/2$ and $D_\alpha = 2e$, see [Bob10, Section 8]. We closely follow the proof therein, but keep track of the α-dependent constants.

Proof. We begin with the first inequality. By homogeneity, we assume $\|X\|_{\Psi_\alpha} = 1$. First let us show that we have

$$g(x) := (\alpha e)^{-1/\alpha}\, e^{x^\alpha} - x \geq 0 \qquad \text{for} \quad x \geq 0. \tag{B.2}$$

Note that g is continuous on $[0,\infty)$ and differentiable on $(0,\infty)$ with $g(0) > 0$ and $g(x) \to \infty$ as $x \to \infty$. Therefore, it suffices to find the critical points. We can rewrite the condition $g'(x) = 0$ as $e^y y = y^{1/\alpha} (\alpha e)^{1/\alpha}$, setting $y := x^\alpha$. From this representation it can be seen that there can be at most two points x_0 and x_1 satisfying this condition. One of these points is $x_\alpha := \alpha^{-1/\alpha}$, which satisfies $g(x_\alpha) = 0$. A short calculation shows that $g''(x_\alpha) = \alpha^{1/\alpha+1} > 0$, so that x_α is a global minimum, from which $g \geq 0$ follows.

Next, from this we can infer for all $p \geq 1$ and $\alpha > 0$

$$x^p \leq \left(\frac{p}{\alpha e}\right)^{p/\alpha} e^{x^\alpha}.$$

Indeed, by a transformation $y = x^p$ and the change $\widetilde{\alpha} = \frac{\alpha}{p}$ this is just an application of (B.2). Consequently, for any $p \geq 1$ it holds

$$\mathbb{E}|X|^p \leq \left(\frac{p}{\alpha e}\right)^{p/\alpha} \mathbb{E}\exp\left(|X|^\alpha\right) \leq 2\left(\frac{p}{\alpha e}\right)^{p/\alpha} \leq 2^p \left(\frac{p}{\alpha e}\right)^{p/\alpha},$$

i. e.

$$\|X\|_p \leq 2(\alpha e)^{-1/\alpha} p^{1/\alpha}.$$

For the second inequality, assume $\sup_{p\geq 1} \frac{\|X\|_p}{p^{1/\alpha}} = 1$, again by homogeneity. First, we need to extend the the supremum to $p \in [\alpha, \infty)$, which can be done as follows. For any $p \in [\alpha, 1)$ we have

$$\frac{\|X\|_p}{p^{1/\alpha}} \leq \frac{\|X\|_1}{p^{1/\alpha}} \leq \frac{1}{p^{1/\alpha}} \leq \frac{1}{\alpha^{1/\alpha}} \Longrightarrow \sup_{p\geq\alpha} \frac{\|X\|_p}{p^{1/\alpha}} \leq \frac{1}{\alpha^{1/\alpha}}.$$

Now, by Taylor's expansion and using the inequality $n^n \leq e^n n!$ we obtain

$$\mathbb{E}\exp\left(\frac{|X|^\alpha}{t^\alpha}\right) = 1 + \sum_{n=1}^{\infty} \frac{\mathbb{E}|X|^{\alpha n}}{t^{\alpha n} n!} \leq 1 + \sum_{n=1}^{\infty} \frac{n^n}{n! t^{\alpha n}} \leq 1 + \sum_{n=1}^{\infty} \left(\frac{e}{t^\alpha}\right)^n = \frac{1}{1 - e/t^\alpha}.$$

For $t = (2e)^{1/\alpha}$ this is less or equal to 2, so that

$$\|X\|_{\Psi_\alpha} \leq (2e)^{1/\alpha} \sup_{p\geq 1} \frac{\|X\|_p}{p^{1/\alpha}}.$$

□

Lemma B.3. *For any* $0 < \alpha < 1$ *and any random variables* X,Y *it holds*

$$\|X + Y\|_{\Psi_\alpha} \leq 2^{1/\alpha}\left(\|X\|_{\Psi_\alpha} + \|Y\|_{\Psi_\alpha}\right).$$

Proof. Let $K \coloneqq \|X\|_{\Psi_\alpha}$ and $L \coloneqq \|Y\|_{\Psi_\alpha}$ and define $t \coloneqq 2^{1/\alpha}(K + L)$. We have

$$\begin{aligned}
\mathbb{E}\exp\left(\frac{|X+Y|^\alpha}{t^\alpha}\right) &\leq \mathbb{E}\exp\left(\frac{(|X|+|Y|)^\alpha}{t^\alpha}\right) \leq \mathbb{E}\exp\left(\frac{|X|^\alpha + |Y|^\alpha}{2(K+L)^\alpha}\right) \\
&\leq \mathbb{E}\exp\left(\frac{|X|^\alpha}{2K^\alpha}\right)\exp\left(\frac{|Y|^\alpha}{2L^\alpha}\right) \\
&\leq \frac{1}{2}\,\mathbb{E}\exp\left(\frac{|X|^\alpha}{K^\alpha}\right) + \frac{1}{2}\,\mathbb{E}\exp\left(\frac{|Y|^\alpha}{L^\alpha}\right) \leq 2.
\end{aligned}$$

Here, the second step follows from the inequality $(x + y)^\alpha \leq x^\alpha + y^\alpha$ valid for all x,$y \geq 0$ and $\alpha \in [0,1]$, and the fourth one is an application of Young's inequality

$ab \leq a^2/2 + b^2/2$ for all $a,b \geq 0$. □

Lemma B.4. *Let $0 < \alpha < 1$. For all random variables X we have*

$$\|\mathbb{E}\,X\|_{\Psi_\alpha} \leq \frac{1}{d_\alpha (\log 2)^{1/\alpha}} \|X\|_{\Psi_\alpha}.$$

Proof. Assuming $\|X\|_{\Psi_\alpha} < \infty$, an application of Lemma B.2 gives

$$\|\mathbb{E}\,X\|_{\Psi_\alpha} = \frac{|\mathbb{E}\,X|}{(\log 2)^{1/\alpha}} \leq \frac{\|X\|_1}{(\log 2)^{1/\alpha}} \leq \frac{1}{d_\alpha (\log 2)^{1/\alpha}} \|X\|_{\Psi_\alpha}.$$

□

We can readily infer the following corollary from the last two lemmas.

Corollary B.5. *For any $0 < \alpha < 1$ and any random variable X it holds*

$$\|X - \mathbb{E}\,X\|_{\Psi_\alpha} \leq 2^{1/\alpha} \left(1 + (d_\alpha \log 2)^{-1/\alpha}\right) \|X\|_{\Psi_\alpha}.$$

APPENDIX C

LSIs and difference operators

To conclude this thesis, we discuss the LSI property (2.3) for different choices of difference operators Γ. Here, we always assume that the probability measure μ is defined on a product of Polish spaces $\mathcal{Y} = \otimes_{i=1}^n \mathcal{X}_i$ (which itself is a Polish space) and equipped with the product Borel σ-algebra $\mathcal{A} = \mathcal{B}(\otimes_{i=1}^n \mathcal{X}_i)$.

As stated in Chapter 2, we can use the disintegration theorem on Polish spaces to define the difference operators $\mathfrak{h}$ and $\mathfrak{d}$. For finite spaces, $\mu(\cdot \mid \overline{x}_i)$ is just the ordinary conditional probability as used in the definition of the difference operator $\mathfrak{d}$. The setting of Polish spaces is clearly more general than the one of finite spin systems. However, the next proposition shows that the $\mathfrak{d}-$LSI property in fact enforces the underlying space to be finite. More precisely, in this case we say that μ has finite support if there is no sequence of sets $A_k \in \mathcal{A}$ with $\mu(A_k) > 0$ for any k and $\mu(A_k) \to 0$.

Proposition C.1. *Let $\mathcal{Y} = \otimes_{i=1}^n \mathcal{X}_i$ be a product of Polish spaces, and μ be a probability measure on $\mathcal{Y}$. If μ satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ for some $\sigma^2 < \infty$, then μ has finite support. Moreover, if μ is a product probability measure, then μ satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ if and only if μ has finite support.*

Proof. First assume μ does not have finite support, i. e. there exists a sequence $A_k \in \mathcal{A}$ with $\mu(A_k) \to 0$. Choosing $f_k := \mathbb{1}_{A_k} \in L^\infty(\mu)$ and assuming a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ holds, we obtain

$$\mu(A_k)\log(1/\mu(A_k)) = \mathrm{Ent}_\mu(f_k^2) \le 2\sigma^2 \int (\mathfrak{d} f_k)^2 d\mu = 2\sigma^2 \mu(A_k)(1-\mu(A_k)). \quad \text{(C.1)}$$

This easily leads to a contradiction as $k \to \infty$.

On the other hand, let μ be a product probability measure with finite support. By tensorization, it suffices to consider $n = 1$, and we may moreover assume $\mathcal{Y}$ to have finitely many elements only. Then, by [BT06, Remark 6.6], μ satisfies a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$ with $\sigma^2 \le C \log(1/\min_{y:\mu(y)>0} \mu(y))$, which finishes the proof. □

In fact, Proposition C.1 can be adapted to the difference operator $\mathfrak{h}^+$ as well. To see this, note that (C.1) can easily be rewritten for the difference operator $\mathfrak{h}^+$ (with only minor changes) and $\int |\mathfrak{d} f|^2 d\mu \le \int |\mathfrak{h}^+ f|^2 d\mu$. In particular, the $\mathfrak{d}$- and $\mathfrak{h}^+$-LSI properties are not essentially different.

The situation drastically changes if we consider $\mathfrak{h}-$LSIs instead. Here, a sufficient condition for the $\mathfrak{h}-$LSI property to hold is that the measure μ satisfies an approximate tensorization (AT) property. As a consequence, for product probability measures, satisfying an $\mathfrak{h}$-LSI is in fact a universal property.

Theorem C.2. *Let $\mathcal{Y} = \otimes_{i=1}^n \mathcal{X}_i$ be a product of Polish spaces, and μ be a probability measure on $\mathcal{Y}$. If μ satisfies an approximate tensorization property*

$$\mathrm{Ent}_\mu(f^2) \le C \sum_{i=1}^n \int \mathrm{Ent}_{\mu(\cdot|\overline{x}_i)}(f^2(\overline{x}_i, \cdot)) d\overline{\mu}_i(\overline{x}_i),$$

then μ also satisfies an $\mathfrak{h}-\mathrm{LSI}(C)$. In particular, any product probability measure satisfies an $\mathfrak{h}-\mathrm{LSI}(1)$.

For product measures, the theorem might be compared to the Efron–Stein inequality (see e. g. [ES81; Ste86]) which establishes the tensorization property for the variance, and can be regarded as a universal Poincaré inequality with respect to $\mathfrak{d}$ (see [BGS19] for such an interpretation). However, Theorem C.2 does not imply the Efron–Stein inequality due to the usage of $\mathfrak{h}$ instead of $\mathfrak{d}$. Unfortunately, as Proposition C.1 demonstrates, there is no "entropy version" of the Efron–Stein inequality of the form $\mathrm{Ent}_\mu(f^2) \le C\,\mathbb{E}_\mu |\mathfrak{d} f|^2$ valid for product probability measure μ and some universal constant C.

Unfortunately, it seems impossible to use the entropy method based on $\mathfrak{h}-$LSI. More precisely, Theorem C.2 cannot be used to estimate the growth of L^p norms as in the setting of a $\mathfrak{d}-\mathrm{LSI}(\sigma^2)$. Indeed, it is impossible to prove the required moment inequalities

$$\|f - \mathbb{E}\, f\|_p \le (\sigma^2 p)^{1/2} \|\mathfrak{h} f\|_p \tag{C.2}$$

under an $\mathfrak{h}-\mathrm{LSI}(\sigma^2)$. For example, the measure $\mu_q = q\delta_1 + (1-q)\delta_0$ satisfies $\mathfrak{h}-\mathrm{LSI}(\sigma_q^2)$ with $\sigma_q^2 \sim q(1-q)\log(1/q)$ (for $q \to 0$), so that (C.2) would imply for $f(x) = x$ an upper bound on the Orlicz norm associated to $\varPsi_2(x) = e^{x^2} - 1$

$$\|f - \mathbb{E}\, f\|_{\varPsi_2} \le 2e \sup_{q \ge 1} \frac{\|f - \mathbb{E}\, f\|_q}{q^{1/2}} \le 4e\sigma_p.$$

However, a simple calculation shows that $\mathbb{E} \exp\big(\frac{(f-\mathbb{E} f)^2}{16 e^2 \sigma_p^2}\big) \to \infty$ as $p \to 0$.

The approximate tensorization property in Theorem C.2 is interesting in its own right, but it is not yet well-studied. We have seen sufficient conditions in Appendix A. Similar results have been derived in [CMT15], which can be applied in discrete and continuous settings. For example, if one considers a measure of the form

$$\mu(x) = Z^{-1} \prod_{i=1}^n \mu_{0,i}(x_i) \exp\Big(\sum_{i,j} J_{ij} w_{ij}(x_i, x_j) \Big)$$

for some countable spaces $\varOmega_i$, $x_i \in \varOmega_i$, measures $\mu_{0,i}$ on $\varOmega_i$ and bounded functions

w_{ij}, under certain technical conditions μ satisfies an approximate tensorization property, which is independent of any functional inequality for $\mu_{0,i}$.

On the other hand, the AT(C) property requires some certain weak dependence conditions. For example, the push-forward of a random permutation π of $[n]$ to $\mathbb{N}^n$ cannot satisfy an approximate tensorization property. It is an interesting question to find necessary and sufficient conditions for the approximate tensorization property to hold.

Now we prove Theorem C.2. Here we include a simplified proof, for the original proof we refer to [GSS18, Theorem 5.2].

Proof of Theorem C.2. Our first step is to that for any bounded, measurable function f the inequality

$$\mathrm{Ent}_\mu(f^2) \le 2 \sup_{x,y} (f(x) - f(y))^2 \tag{C.3}$$

holds. From $\log(x) + 1 \le x$ for all $x > 0$ we can deduce

$$x^2 - x \log x - x \ge 0 \quad \text{for all} \quad x \ge 0. \tag{C.4}$$

Let $f \ge 0$ be a measurable function satisfying $\int f d\mu = 1$. Integrating (C.4) with respect to μ yields

$$\int f^2 d\mu - 1 \ge \int f \log f d\mu,$$

i. e. by using the definition and the homogeneity of the entropy we obtain for any measurable function f

$$\int f^2 d\mu \, \mathrm{Ent}_\mu(f^2) \le \mathrm{Var}_\mu(f^2).$$

Moreover, we have the upper bound

$$\begin{aligned} \mathrm{Var}_\mu(f^2) &= \frac{1}{2} \iint (f(x) - f(y))^2 (f(x) + f(y))^2 d\mu(x) d\mu(y) \\ &\le 2 \sup_{x,y} (f(x) - f(y))^2 \int f^2 d\mu. \end{aligned}$$

This proves (C.3) and the case $n = 1$.

For arbitrary n, the proof is now easily completed. Assume that $f \in L^\infty(\mu)$, i. e. $\overline{\mu}_i(\overline{x}_i)$-a.s. we have $f(\overline{x}_i, \cdot) \in L^\infty(\mu(\cdot \mid \overline{x}_i))$. For these $\overline{x}_i$, the $n = 1$ case yields

$$\mathrm{Ent}_{\mu(\cdot \mid \overline{x}_i)}(f^2(\overline{x}_i, \cdot)) \le 2 \sup_{y_i', y_i''} \big(f(\overline{x}_i, y_i') - f(\overline{x}_i, y_i'') \big)^2.$$

Plugging this into the assumption leads to

$$\mathrm{Ent}_\mu(f^2) \le 2C \sum_{i=1}^n \int \sup_{y_i', y_i''} \big(f(\overline{x}_i, y_i') - f(\overline{x}_i, y_i'') \big)^2 d\overline{\mu}_i(\overline{x}_i) = 2C \int |\mathfrak{h} f|^2 d\mu.$$

As for the second part, it is a classical fact that product measures satisfy the tensorization property (i. e. AT(1)), see for example [Led01, Proposition 5.6], [BBLM05, Theorem 4.10] or [Han16, Theorem 3.14]. Moreover, in this case, the assumption that $\mathcal{Y}$ is a product of Polish spaces can be dropped by simply defining $\mu(\cdot \mid \overline{x}_i) = \mu_i$. $\square$

Open questions

Question 1. The approximate tensorization of entropy property for weakly dependent spin systems has two major drawbacks. First off, it relies on Lemma A.1 and thus on the comparability of the relative entropy and the total variation distance. However, these can only be compared in finite spaces, severely limiting the applicability of Theorem A.2. There are two other works proving approximate tensorization of entropy:

- in [Mar13] the author has succeeded in proving it for measures of the form $\exp(-V(x))dx$ in $\mathbb{R}^n$ under technical conditions on V, but it relied on properties of $\mathbb{R}^n$,
- [CMT15] used weak dependence condition allowing for the approximate tensorization for countable spaces as well, but it yields sub-optimal conditions in the Curie–Weiss model on $\{-1,+1\}^n$, whereas Theorem A.2 can be applied in the full regime.

The second problem arises from the fact that the interdependence matrix J always assumes the worst possible configuration x. However, this does not take into account the fact that many models tend to be heavily concentrated on “typical configurations”, and thus weaker conditions should suffice to establish an entropy tensorization property.

Consequently, there are (at least) two natural questions. How can one relax the condition on the “worst possible configuration” to account for model-specific concentration properties? This could lead to the approximate tensorization of entropy property in larger subsets of the parameter domain, maybe optimal in more models (e. g. in the exponential random graph model), and thus by the same approach as taken in this thesis to concentration of measure. And secondly, is it possible to remove the condition of $\mathcal{X}$ being finite? As mentioned above, another method was employed in [CMT15], but the generalization should yield the optimal range of β in the arguably simplest non-product model - the Curie–Weiss model.

Question 2. Theorem A.2 is only applicable for a “true” Glauber dynamic, i. e. if μ is a spin system on $\mathcal{X}^{\mathcal{I}}$, then only one $i \in \mathcal{I}$ is chosen uniformly and updated with the conditional probability. However, there are many interesting models in which the Glauber dynamic is actually trivial.

In the first example, take $\mathcal{X} = [n]$ and $\mathcal{I} = [n]$ and let μ be a spin system on $\mathcal{X}^{\mathcal{I}}$ which is given by the push-forward of the map $S_n \ni \sigma \mapsto (\sigma(1), \ldots, \sigma(n))$ under the uniform distribution on the symmetric group S_n. It is clear that μ is supported on configurations with n distinct numbers, and thus we have $\mu(\cdot \mid \overline{x}_i) = \delta_y$ for some uniquely determined $y \in [n]$. Thus, the Glauber dynamic does not leave its starting state. However, if we change the choice of the index, the new dynamics becomes very interesting. Instead of updating a single $i \in \mathcal{I}$, we can update a pair $(i,j) \in \mathcal{I}^2$: In this case, we have $\mu(\cdot \mid \overline{x}_{ij}) = \frac{1}{2}\delta_{(y,z)} + \frac{1}{2}\delta_{(z,y)}$, and so the dynamics is the (lazy, if (i,i) is allowed) random walk on S_n (by identification) generated by the transpositions.

A second example is the random walk on the set of all d-regular graphs on n vertices. Again, if we choose a single edge for an update, the Glauber dynamic remains in its current state, as all the vertices need to have degree d. However, if we choose four edges $e_1, \ldots, e_4$ and consider the conditional probability, a short calculation shows that there are only two possibilities in each step - either the original configuration remains, or a simple switching is performed.

In light of these two examples, it is a natural question how to generalize the approach in Theorem A.2 to such “Glauber-type” dynamics, and provide sufficient conditions for the system to be weakly dependent. Note that there is an additional difficulty not present in the one-site Glauber dynamic. If we denote by $\mathcal{J} \subset \mathcal{P}(\mathcal{I})$ the set of “allowed updates”, then for $J_1, J_2 \in \mathcal{J}$ with $J_1 \neq J_2$ we might encounter $J_1 \cap J_2 \neq \emptyset$. A useful source of inspiration might be [Mar04], where similar thoughts have been carried out in the $\mathbb{R}^n$ setting. However, [Mar04] does not prove a logarithmic Sobolev, but a Talagrand inequality.

Question 3. Another famous route to obtaining concentration inequalities is by using the *method of exchangeable pairs* pioneered by Chatterjee [Cha05; Cha07] (at least in the framework of measure concentration; the concept was introduced by Stein to prove distributional limits with explicit rates of convergence in the seventies). It is applicable in a wide variety of different settings and provides sub-Gaussian and Bernstein-type inequalities, see [Cha07, Theorem 1.5] and all the examples discussed in the two works. The main equality which allows for various identities and concentration of measure statements is that for any exchangeable pair (X,X'), any function $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $F(x,y) = -F(y,x)$ and any $h : \mathcal{X} \to \mathbb{R}$ it holds

$$\mathbb{E}\, h(X)f(X) = \frac{1}{2}\,\mathbb{E}\left((h(X) - h(X'))F(X,X')\right) = \mathbb{E}\left((h(X) - h(X'))_+ F(X,X')\right),$$

where $f(X) \coloneqq \mathbb{E}(F(X,X') \mid X)$. It is very tempting to set $h(X) \coloneqq \exp(f(X))$ and understand the left-hand side as the covariance of $\exp(f(X))$ and $f(X)$ (since $f(X)$ is centered as shown in [Cha07, Theorem 1.5]), which itself (by Jensen's inequality) is at least as large as the entropy of $\exp(f(X))$. There seems to be a connection between these two concepts, but one cannot set

$$\mathit{\Gamma} f(X)^2 \coloneqq \mathbb{E}\left[(f(X) - f(X'))_+ F(X,X') \mid X\right],$$

as this does not satisfy the conditions of a difference operator. Actually, the set of all functions $f : \mathcal{X} \to \mathbb{R}$ possessing an anti-symmetric function F does not contain the constants, as all such functions satisfy $\mathbb{E} f(X) = 0$. However, it is clear that $\Gamma(\lambda f)^2 = \lambda^2 \Gamma(f)^2$, which might be sufficient to establish a connection to [BG99].

Question 4. In Chapter 3 we have shown modified LSIs for the symmetric group and established deviation and concentration inequalities using the entropy method. Furthermore, it was possible to derive a weak form of the convex distance inequality as proven in [Tal95] by mimicking the approach in [BLM03]. In [BLM09] the authors have established the strong form of Talagrand's convex distance inequality for independent random variables using the entropy method as well. The question arises whether it is also possible to prove a stronger functional inequality than in [GQ03] for random permutations to obtain the convex distance inequality in this setting as well. Additionally, this could shed some light on the question for which models one can expect an analogue of a convex distance inequality.

Question 5. Concerning Chapter 4 and the deviation inequalities for suprema type functionals, it is still an open question to provide concentration inequalities for $f(X) = \sup_{t \in \mathcal{T}} \|\sum_I t_I X_I\|$. The result in [Tal96b] is a "true" concentration inequality, and not a deviation inequality for the upper tail. However, all the possible attempts at generalizing this result produce deviation inequalities (see [BBLM05], [Ada15], [KZ18] and the results in Chapter 4). The sole two-sided concentration result is [Ada15, Theorem 2.10] under rather restrictive conditions (which – maybe somewhat surprisingly – does not include the Rademacher case).

Question 6. In Chapter 5 we have explicitly calculated all possible norms $\|A\|_{\mathcal{I}}$ in the case $d = 2$, i. e. for a matrix A. Is it possible to extend this to $d > 2$, for example in the easiest case $q = 1$? Clearly, there are many more cases to consider, but all norms should be the maximal ℓ^2 norms of any $d - k$-dimensional sub-tensor, as these quantities appear in [KL15].

Question 7. This question might be the most challenging one, but perhaps the most interesting as well. There have been many approaches to tackle the triangle problem in Erdös–Rényi random graphs, and concentration properties for general polynomials do not seem to catch the behavior for values of p very close to the threshold n^{-1}. Starting from the papers [KV00; Vu02] concentration properties of polynomials have been considered in many article, which we have partly mentioned in this thesis. [AW15] provides exponential inequalities with an optimal exponent, but the range was limited to $p \geq n^{-1/4}$ (up to some logarithmic terms). From the perspective of concentration of measure, it would be desirable to find general exponential inequalities for polynomials, which retrieve the well-known optimal (see [Cha12; DK12]) tail decay in the full range $p \geq n^{-1}$. Furthermore, it is very interesting to establish some kind of connection between multilevel concentration inequalities for polynomials and the Kim–Vu inequalities.

Bibliography

[Ada06] Radosław Adamczak. 'Moment inequalities for U-statistics'. In: *Ann. Probab.* **34.6** (2006), pages 2288–2314. DOI: 10.1214/009117906000000476.

[Ada08] Radosław Adamczak. 'A tail inequality for suprema of unbounded empirical processes with applications to Markov chains'. In: *Electron. J. Probab.* **13** (2008), no. 34, 1000–1034. DOI: 10.1214/EJP.v13-521.

[Ada15] Radosław Adamczak. 'A note on the Hanson–Wright inequality for random vectors with dependencies'. In: *Electron. Commun. Probab.* **20** (2015), no. 72, 13. DOI: 10.1214/ECP.v20-3829.

[ABW17] Radosław Adamczak, Witold Bednorz, and Paweł Wolff. 'Moment estimates implied by modified log-Sobolev inequalities'. In: *ESAIM Probab. Stat.* **21** (2017), pages 467–494. DOI: 10.1051/ps/2016030.

[AKPS19] Radosław Adamczak, Michał Kotowski, Bartłomiej Polaczyk, and Michał Strzelecki. 'A note on concentration for polynomials in the Ising model'. In: *Electron. J. Probab.* **24** (2019), no. 42, 1–22. DOI: 10.1214/19-EJP280.

[AL12] Radosław Adamczak and Rafał Latała. 'Tail and moment estimates for chaoses generated by symmetric random variables with logarithmically concave tails'. In: *Ann. Inst. Henri Poincaré Probab. Stat.* **48.4** (2012), pages 1103–1136. DOI: 10.1214/11-AIHP441.

[ALM18] Radosław Adamczak, Rafał Latała, and Rafał Meller. 'Hanson–Wright inequality in Banach spaces'. In: *arXiv preprint* (2018). arXiv: 1811.00353.

[AW15] Radosław Adamczak and Paweł Wolff. 'Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order'. In: *Probab. Theory Related Fields* **162.3-4** (2015), pages 531–586. DOI: 10.1007/s00440-014-0579-3.

[AS94] Shigeki Aida and Daniel W. Stroock. 'Moment estimates derived from Poincaré and logarithmic Sobolev inequalities'. In: *Math. Res. Lett.* **1.1** (1994), pages 75–86. DOI: 10.4310/MRL.1994.v1.n1.a9.

[AG13] Luigi Ambrosio and Nicola Gigli. 'A user's guide to optimal transport'. In: *Modelling and optimisation of flows on networks.* Volume 2062. Lecture Notes in Mathematics. Springer, Berlin-Heidelberg, 2013, pages 1–155. DOI: 10.1007/978-3-642-32160-3_1.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Second edition. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pages x+334. ISBN: 978-3-7643-8721-1.

[Azu67] Kazuoki Azuma. 'Weighted sums of certain dependent random variables'. In: *Tôhoku Math. J. (2)* **19** (1967), pages 357–367. DOI: 10.2748/tmj/1178243286.

[Ben62] George Bennett. 'Probability inequalities for the sum of independent random variables'. In: *J. Amer. Statist. Assoc.* **57.297** (1962), pages 33–45. DOI: 10.2307/2282438.

[Ber24] Sergei N. Bernstein. 'On a modification of Chebyshev's inequality and of the error formula of Laplace'. In: *Annals Science Institute Sav. Ukraine, Sect. Math. I* (1924). Russian, French summary.

[BBS11] Shankar Bhamidi, Guy Bresler, and Allan Sly. 'Mixing time of exponential random graphs'. In: *Ann. Appl. Probab.* **21.6** (2011), pages 2146–2170. DOI: 10.1214/10-AAP740.

[Bil68] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, Inc., New York-London-Sydney, 1968, pages xii+253. ISBN: 978-0-471-19745-4.

[Bob00] Sergey G. Bobkov. 'Remarks on the growth of L^p-norms of polynomials'. In: *Geometric Aspects of Functional Analysis*. Volume 1745. Lecture Notes in Mathematics. Springer, Berlin, 2000, pages 27–35. DOI: 10.1007/BFb0107206.

[Bob10] Sergey G. Bobkov. 'The growth of L^p-norms in presence of logarithmic Sobolev inequalities'. In: *Vestnik Syktyvkar Univ.* **11.2** (2010), pages 92–111. URL: http://www-users.math.umn.edu/~bobko001/papers/2010_Vestnik_Growth.under.log-Sobolev.pdf.

[BCG17] Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. 'Second-order concentration on the sphere'. In: *Commun. Contemp. Math.* **19.5**, 1650058 (2017). DOI: 10.1142/S0219199716500589.

[BG99] Sergey G. Bobkov and Friedrich Götze. 'Exponential integrability and transportation cost related to logarithmic Sobolev inequalities'. In: *J. Funct. Anal.* **163.1** (1999), pages 1–28. DOI: 10.1006/jfan.1998.3326.

[BG07] Sergey G. Bobkov and Friedrich Götze. 'Concentration inequalities and limit theorems for randomized sums'. In: *Probab. Theory Related Fields* **137.1-2** (2007), pages 49–81. DOI: 10.1007/s00440-006-0500-9.

[BGS19] Sergey G. Bobkov, Friedrich Götze, and Holger Sambale. 'Higher order concentration of measure'. In: *Commun. Contemp. Math.* **21.03**, 1850043 (2019). DOI: 10.1142/S0219199718500438.

[BL97] Sergey G. Bobkov and Michel Ledoux. 'Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution'. In: *Probab. Theory Related Fields* **107.3** (1997), pages 383–400. DOI: 10.1007/s004400050090.

[BT06] Sergey G. Bobkov and Prasad Tetali. 'Modified logarithmic Sobolev inequalities in discrete settings'. In: *J. Theoret. Probab.* **19.2** (2006), pages 289–336. DOI: 10.1007/s10959-006-0016-3.

[Bor75] Christer Borell. 'The Brunn-Minkowski inequality in Gauss space'. In: *Invent. Math.* **30.2** (1975), pages 207–216. DOI: 10.1007/BF01425510.

[BBLM05] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. 'Moment inequalities for functions of independent random variables'. In: *Ann. Probab.* **33.2** (2005), pages 514–560. DOI: 10.1214/009117904000000856.

[BLM00] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 'A sharp concentration inequality with applications'. In: *Random Structures Algorithms* **16.3** (2000), pages 277–292. DOI: 10.1002/(SICI)1098-2418(200005)16:3<277::AID-RSA4>3.0.CO;2-1.

[BLM03] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 'Concentration inequalities using the entropy method'. In: *Ann. Probab.* **31.3** (2003), pages 1583–1614. DOI: 10.1214/aop/1055425791.

[BLM09] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 'On concentration of self-bounding functions'. In: *Electron. J. Probab.* **14** (2009), no. 64, 1884–1899. DOI: 10.1214/EJP.v14-690.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013, pages x+481. ISBN: 978-0-19953-525-5.

[Bou91] Jean Bourgain. 'On the distribution of polynomials on high-dimensional convex sets'. In: *Geometric Aspects of Functional Analysis*. Volume 1469. Lecture Notes in Mathematics. Springer, Berlin-Heidelberg, 1991, pages 127–137. DOI: 10.1007/BFb0089219.

[Bou02] Olivier Bousquet. 'A Bennett concentration inequality and its application to suprema of empirical processes'. In: *C. R. Math. Acad. Sci. Paris* **334.6** (2002), pages 495–500. DOI: 10.1016/S1631-073X(02)02292-6.

[CTP19] Joshua Cape, Minh Tang, and Carey E. Priebe. 'The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics'. In: *Ann. Statist.* **47.5** (2019), pages 2405–2439. DOI: 10.1214/18-AOS1752.

[CMT15] Pietro Caputo, Georg Menz, and Prasad Tetali. 'Approximate tensorization of entropy at high temperature'. In: *Ann. Fac. Sci. Toulouse Math. (6)* **24.4** (2015). DOI: 10.5802/afst.1460.

[CKSS72] Leonard Carlitz, David C. Kurtz, Richard Scoville, and Olaf P. Stackelberg. 'Asymptotic properties of Eulerian numbers'. In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **23** (1972), pages 47–54. DOI: 10.1007/BF00536689.

[Cha04] Djalil Chafaï. 'Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities'. In: *J. Math. Kyoto Univ.* **44.2** (2004), pages 325–363. DOI: 10.1215/kjm/1250283556.

[Cha05] Sourav Chatterjee. 'Concentration inequalities with exchangeable pairs'. PhD thesis. Stanford University, 2005. URL: https://arxiv.org/abs/math/0507526.

[Cha07] Sourav Chatterjee. 'Stein's method for concentration inequalities'. In: *Probab. Theory Related Fields* **138.1-2** (2007), pages 305–321. DOI: 10.1007/s00440-006-0029-y.

[Cha12] Sourav Chatterjee. 'The missing log in large deviations for triangle counts'. In: *Random Structures Algorithms* **40.4** (2012), pages 437–451. DOI: 10.1002/rsa.20381.

[Cha16] Sourav Chatterjee. 'An introduction to large deviations for random graphs'. In: *Bull. Amer. Math. Soc. (N.S.)* **53.4** (2016), pages 617–642. DOI: 10.1090/bull/1539.

[Cha17] Sourav Chatterjee. *Large deviations for random graphs*. Volume 2197. Lecture Notes in Mathematics. Springer, Cham, 2017, pages xi+167. ISBN: 978-3-319-65815-5.

[CD13] Sourav Chatterjee and Persi Diaconis. 'Estimating and understanding exponential random graph models'. In: *Ann. Statist.* **41.5** (2013), pages 2428–2461. DOI: 10.1214/13-AOS1155.

[CD17] Sourav Chatterjee and Persi Diaconis. 'A central limit theorem for a new statistic on permutations'. In: *Indian J. Pure Appl. Math.* **48.4** (2017), pages 561–573. DOI: 10.1007/s13226-017-0246-3.

[CY18] Xiaohui Chen and Yun Yang. 'Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data'. In: *arXiv preprint* (2018). arXiv: 1810.11180.

[Che81] Herman Chernoff. 'A note on an inequality involving the normal distribution'. In: *Ann. Probab.* **9.3** (1981), pages 533–535. DOI: 10.1214/aop/1176994428.

[CIS76] Boris S. Cirel'son, Ildar A. Ibragimov, and Vladimir N. Sudakov. 'Norms of Gaussian sample functions'. In: *Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975)*. Springer, Berlin-Heidelberg, 1976, 20–41. Lecture Notes in Mathematics, Vol. 550. DOI: 10.1007/BFb0077482.

[Cla09] Lane Clark. 'Limit distribution of ascent, descent or excedance length sums of permutations'. In: *Appl. Anal. Discrete Math.* **3.2** (2009), pages 303–309. DOI: 10.2298/AADM0902303C.

[Cra33] Cecil C. Craig. 'On the Tchebychef inequality of Bernstein'. In: *The Annals of Mathematical Statistics* **4.2** (1933), pages 94–102. DOI: 10.1214/aoms/1177732803.

[DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. 'Concentration of Multilinear Functions of the Ising Model with Applications to Network Data'. In: *arXiv preprint* (2017). arXiv: 1710.04170.

[DM78] Claude Dellacherie and Paul-André Meyer. *Probabilities and potential.* Volume 29. North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam-New York, 1978, pages viii+189. ISBN: 978-0-72040-701-3.

[DK12] Bobby DeMarco and Jeff Kahn. 'Upper tails for triangles'. In: *Random Structures Algorithms* **40.4** (2012), pages 452–459. DOI: 10.1002/rsa.20382.

[DZ10] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications.* Volume 38. Stochastic Modelling and Applied Probability. Corrected reprint of the second (1998) edition. Springer-Verlag, Berlin, 2010, pages xvi+396. ISBN: 978-3-642-03310-0.

[DEY19] Ryan DeMuse, Terry Easlick, and Mei Yin. 'Mixing time of vertex-weighted exponential random graphs'. In: *J. Comput. Appl. Math.* **362** (2019), pages 443–459. DOI: 10.1016/j.cam.2018.07.038.

[Dia88] Persi Diaconis. *Group representations in probability and statistics.* Volume 11. Institute of Mathematical Statistics Lecture Notes—Monograph Series. Institute of Mathematical Statistics, Hayward, CA, 1988, pages vi+198. ISBN: 0-940600-14-5.

[DS96] Persi Diaconis and Laurent Saloff-Coste. 'Logarithmic Sobolev inequalities for finite Markov chains'. In: *Ann. Appl. Probab.* **6.3** (1996), pages 695–750. DOI: 10.1214/aoap/1034968224.

[ES81] Bradley Efron and Charles M. Stein. 'The jackknife estimate of variance'. In: *Ann. Statist.* **9.3** (1981), pages 586–596. DOI: 10.1214/aos/1176345462.

[EL10] Peter Eichelsbacher and Matthias Löwe. 'Stein's method for dependent random variables occurring in statistical mechanics'. In: *Electron. J. Probab.* **15** (2010), no. 30, 962–988. DOI: 10.1214/EJP.v15-777.

[Ell06] Richard S. Ellis. *Entropy, large deviations, and statistical mechanics.* Classics in Mathematics. Reprint of the 1985 original. Springer-Verlag, Berlin, 2006, pages xiv+364. ISBN: 978-3-540-29059-9.

[EN78] Richard S. Ellis and Charles M. Newman. 'Limit theorems for sums of dependent random variables occurring in statistical mechanics'. In: *Z. Wahrsch. Verw. Gebiete* **44.2** (1978), pages 117–139. DOI: 10.1007/BF00533049.

[EYY12] László Erdős, Horng-Tzer Yau, and Jun Yin. 'Bulk universality for generalized Wigner matrices'. In: *Probab. Theory Related Fields* **154.1-2** (2012), pages 341–407. DOI: 10.1007/s00440-011-0390-3.

[FS86] Ove Frank and David Strauss. 'Markov graphs'. In: *J. Amer. Statist. Assoc.* **81.395** (1986), pages 832–842. DOI: 10.2307/2289017.

[GQ03] Fuqing Gao and Jeremy Quastel. 'Exponential decay of entropy in the random transposition and Bernoulli-Laplace models'. In: *Ann. Appl. Probab.* **13.4** (2003), pages 1591–1600. DOI: 10.1214/aoap/1069786512.

[GL13] Sara van de Geer and Johannes Lederer. 'The Bernstein-Orlicz norm and deviation inequalities'. In: *Probab. Theory Related Fields* **157.1-2** (2013), pages 225–250. DOI: 10.1007/s00440-012-0455-y.

[GLP18] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. 'Concentration inequalities for polynomials of contracting Ising models'. In: *Electron. Commun. Probab.* **23** (2018), no. 76, 12. DOI: 10.1214/18-ECP173.

[GLZ00] Evarist Giné, Rafał Latała, and Joel Zinn. 'Exponential and moment inequalities for U-statistics'. In: *High Dimensional Probability II.* Volume 47. Progr. Probab. Birkhäuser Boston, Boston, MA, 2000, pages 13–38. DOI: 10.1007/978-1-4612-1358-1_2.

[GK95] Efim D. Gluskin and Stanisław Kwapień. 'Tail and moment estimates for sums of independent random variables with logarithmically concave tails'. In: *Studia Math.* **114.3** (1995), pages 303–309. DOI: 10.4064/sm-114-3-303-309.

[God55] H. J. Godwin. 'On generalizations of Tchebychef's inequality'. In: *J. Amer. Statist. Assoc.* **50.271** (1955), pages 923–945. DOI: 10.2307/2281177.

[GS19] Friedrich Götze and Holger Sambale. 'Higher order concentration in presence of Poincaré-type inequalities'. In: *High Dimensional Probability VIII.* Volume 74. Progr. Probab. Birkhäuser Verlag, Basel, 2019, pages 55–69. DOI: 10.1007/978-3-030-26391-1.

[GS20] Friedrich Götze and Holger Sambale. 'Second order concentration via logarithmic Sobolev inequalities'. In: *Bernoulli* **26.1** (2020), pages 93–126. DOI: 10.3150/19-BEJ1118.

[GSS18] Friedrich Götze, Holger Sambale, and Arthur Sinulis. 'Concentration inequalities for bounded functionals via log-Sobolev-type inequalities'. In: *arXiv preprint* (2018). arXiv: 1812.01092.

[GSS19a] Friedrich Götze, Holger Sambale, and Arthur Sinulis. 'Concentration inequalities for polynomials in α-sub-exponential random variables'. In: *arXiv preprint* (2019). arXiv: 1903.05964.

[GSS19b] Friedrich Götze, Holger Sambale, and Arthur Sinulis. 'Higher order concentration for functions of weakly dependent random variables'. In: *Electron. J. Probab.* **24** (2019), no. 85, 19 pp. DOI: 10.1214/19-EJP338.

[Gro75] Leonard Gross. 'Logarithmic Sobolev inequalities'. In: *Amer. J. Math.* **97.4** (1975), pages 1061–1083. DOI: 10.2307/2373688.

[Han16] Ramon van Handel. *Probability in High Dimension.* APC 550 Lecture Notes, Princeton University. 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[HW71] David L. Hanson and Farroll T. Wright. 'A bound on tail probabilities for quadratic forms in independent random variables'. In: *Ann. Math. Statist.* **42.3** (1971), pages 1079–1083. DOI: 10.1214/aoms/1177693335.

[HMO97] Paweł Hitczenko, Stephen J. Montgomery-Smith, and Krzysztof Oleszkiewicz. 'Moment inequalities for sums of certain independent symmetric random variables'. In: *Studia Math.* **123.1** (1997), pages 15–42. DOI: 10.4064/sm-123-1-15-42.

[HL81] Paul W. Holland and Samuel Leinhardt. 'An exponential family of probability distributions for directed graphs'. In: *J. Amer. Statist. Assoc.* **76.373** (1981), pages 33–65. DOI: 10.2307/2287037.

[HS87] Richard Holley and Daniel W. Stroock. 'Logarithmic Sobolev inequalities and stochastic Ising models'. In: *J. Statist. Phys.* **46.5-6** (1987), pages 1159–1194. DOI: 10.1007/BF01011161.

[HKZ12] Daniel Hsu, Sham M. Kakade, and Tong Zhang. 'A tail inequality for quadratic forms of subgaussian random vectors'. In: *Electron. Commun. Probab.* **17** (2012), no. 52, 6. DOI: 10.1214/ECP.v17-2079.

[Jer95] Mark Jerrum. 'A very simple algorithm for estimating the number of k-colorings of a low-degree graph'. In: *Random Structures Algorithms* **7.2** (1995), pages 157–165. DOI: 10.1002/rsa.3240070205.

[KLS95] Ravindran Kannan, László Lovász, and Miklós Simonovits. 'Isoperimetric problems for convex bodies and a localization lemma'. In: *Discrete Comput. Geom.* **13.3-4** (1995), pages 541–559. DOI: 10.1007/BF02574061.

[KV00] Jeong H. Kim and Van H. Vu. 'Concentration of multivariate polynomials and its applications'. In: *Combinatorica* **20.3** (2000), pages 417–434. DOI: 10.1007/s004930070014.

[Kla85] Chris A. J. Klaassen. 'On an inequality of Chernoff'. In: *Ann. Probab.* **13.3** (1985), pages 966–974. DOI: 10.1214/aop/1176992917.

[KR05] Thierry Klein and Emannuel Rio. 'Concentration around the mean for maxima of empirical processes'. In: *Ann. Probab.* **33.3** (2005), pages 1060–1077. DOI: 10.1214/009117905000000044.

[KZ18] Yegor Klochkov and Nikita Zhivotovskiy. 'Uniform Hanson–Wright type concentration inequalities for unbounded entries via the entropy method'. In: *arXiv preprint* (2018). arXiv: 1812.03548.

[KL15] Konrad Kolesko and Rafał Latała. 'Moment estimates for chaoses generated by symmetric random variables with logarithmically convex tails'. In: *Statist. Probab. Lett.* **107** (2015), pages 210–214. DOI: 10.1016/j.spl.2015.08.019.

[KC18] Arun K. Kuchibhotla and Abhishek Chakrabortty. 'Moving Beyond Sub-Gaussianity in High-Dimensional Statistics: Applications in Covariance Estimation and Linear Regression'. In: *arXiv preprint* (2018). arXiv: 1804.02605.

[Kül03] Christof Külske. 'Concentration inequalities for functions of Gibbs fields with application to diffraction and random Gibbs measures'. In: *Comm. Math. Phys.* **239.1-2** (2003), pages 29–51. DOI: 10.1007/s00220-003-0841-5.

[Kwa87] Stanisław Kwapień. 'Decoupling inequalities for polynomial chaos'. In: *Ann. Probab.* **15.3** (1987), pages 1062–1071. DOI: 10.1214/aop/1176992081.

[Lat96] Rafał Latała. 'Tail and moment estimates for sums of independent random vectors with logarithmically concave tails'. In: *Studia Math.* **118.3** (1996), pages 301–304. DOI: 10.4064/sm-118-3-301-304.

[Lat99] Rafał Latała. 'Tail and moment estimates for some types of chaos'. In: *Studia Math.* **135.1** (1999), pages 39–53. DOI: 10.4064/sm-135-1-39-53.

[Lat06] Rafał Latała. 'Estimates of moments and tails of Gaussian chaoses'. In: *Ann. Probab.* **34.6** (2006), pages 2315–2331. DOI: 10.1214/009117906000000421.

[LŁ03] Rafał Latała and Rafał Łochowski. 'Moment and tail estimates for multidimensional chaos generated by positive random variables with logarithmically concave tails'. In: *Stochastic inequalities and applications*. Volume 56. Progr. Probab. Birkhäuser, Basel, 2003, pages 77–92. DOI: 10.1007/978-3-0348-8069-5_7.

[Led97] Michel Ledoux. 'On Talagrand's deviation inequalities for product measures'. In: *ESAIM Probab. Statist.* **1** (1997), pages 63–87. DOI: 10.1051/ps:1997103.

[Led01] Michel Ledoux. *The concentration of measure phenomenon.* Volume 89. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001, pages x+181. ISBN: 978-0-8218-3792-4.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Volume 23. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer-Verlag, Berlin, 1991, pages xii+480. ISBN: 978-3-642-20211-7.

[LS93] László Lovász and Miklós Simonovits. 'Random walks in a convex body and an improved volume algorithm'. In: *Random Structures Algorithms* **4.4** (1993), pages 359–412. DOI: 10.1002/rsa.3240040402.

[Mar04] Katalin Marton. 'Measure concentration for Euclidean distance in the case of dependent random variables'. In: *Ann. Probab.* **32.3B** (2004), pages 2526–2544. DOI: 10.1214/009117904000000702.

[Mar13] Katalin Marton. 'An inequality for relative entropy and logarithmic Sobolev inequalities in Euclidean spaces'. In: *J. Funct. Anal.* **264.1** (2013), pages 34–61. DOI: 10.1016/j.jfa.2012.10.001.

[Mar15] Katalin Marton. 'Logarithmic Sobolev inequalities in discrete product spaces: a proof by a transportation cost distance'. In: *arXiv preprint* (2015). arXiv: 1507.02803.

[Mas00] Pascal Massart. 'About the constants in Talagrand's concentration inequalities for empirical processes'. In: *Ann. Probab.* **28.2** (2000), pages 863–884. DOI: 10.1214/aop/1019160263.

[McD89] Colin McDiarmid. 'On the method of bounded differences'. In: *Surveys in combinatorics, 1989 (Norwich, 1989).* Volume 141. London Math. Soc. Lecture Note Ser. Cambridge University Press, Cambridge, 1989, pages 148–188. DOI: 10.1017/CBO9781107359949.008.

[MR06] Colin McDiarmid and Bruce Reed. 'Concentration for self-bounding functions and an inequality of Talagrand'. In: *Random Structures Algorithms* **29.4** (2006), pages 549–557. DOI: 10.1002/rsa.20145.

[Mel16] Rafał Meller. 'Two-sided moment estimates for a class of nonnegative chaoses'. In: *Statist. Probab. Lett.* **119** (2016), pages 213–219. DOI: 10.1016/j.spl.2016.08.005.

[Mel19] Rafał Meller. 'Tail and moment estimates for a class of random chaoses of order two'. In: *Studia Math.* **249.1** (2019), pages 1–32. DOI: 10.4064/sm170819-2-5.

[Mil71] Vitali D. Milman. 'New proof of the theorem of A. Dvoretzky's on intersections of convex bodies'. In: *Funkcional. Anal. i Priložen.* **5.4** (1971), pages 28–37. DOI: 10.1007/BF01086740.

[MS86] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces.* Volume 1200. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1986, pages viii+156. ISBN: 978-3-540-16769-3.

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. 'Noise stability of functions with low influences: invariance and optimality'. In: *Ann. of Math. (2)* **171.1** (2010), pages 295–341. DOI: 10.4007/annals.2010.171.295.

[MW82] Carl E. Mueller and Fred B. Weissler. 'Hypercontractivity for the heat semigroup for ultraspherical polynomials and on the n-sphere'. In: *J. Functional Analysis* **48.2** (1982), pages 252–283. DOI: 10.1016/0022-1236(82)90069-6.

[Muk13a] Sumit Mukherjee. 'Consistent estimation in the two star exponential random graph model'. In: *arXiv preprint* (2013). arXiv: 1310.4526.

[Muk13b] Sumit Mukherjee. 'Phase transition in the two star exponential random graph model'. In: *arXiv preprint* (2013). arXiv: 1310.4164.

[NSU19] Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. 'Bootstrap confidence sets for spectral projectors of sample covariance'. In: *Probab. Theory Related Fields* **174.3-4** (2019), pages 1091–1132. DOI: 10.1007/s00440-018-0877-2.

[Nel73] Edward Nelson. 'The free Markoff field'. In: *J. Functional Analysis* **12.2** (1973), pages 211–227. DOI: 10.1016/0022-1236(73)90025-6.

[ODo14] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, New York, 2014, pages xx+423. ISBN: 978-1-107-03832-5.

[Oli10] Roberto Imbuzeiro Oliveira. 'Sums of random Hermitian matrices and an inequality by Rudelson'. In: *Electron. Commun. Probab.* **15** (2010), pages 203–212. DOI: 10.1214/ECP.v15-1544.

[Pau14] Daniel Paulin. 'The convex distance inequality for dependent random variables, with applications to the stochastic travelling salesman and other problems'. In: *Electron. J. Probab.* **19** (2014), no. 68, 34. DOI: 10.1214/EJP.v19-3261.

[PG99] Víctor H. de la Peña and Evarist Giné. *Decoupling. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.* Probability and its Applications. Springer-Verlag, New York, 1999, pages xvi+392. ISBN: 978-0-387-98616-6.

[PTT19] Joscha Prochno, Christoph Thäle, and Nicola Turchi. 'The isotropic constant of random polytopes with vertices on convex surfaces'. In: *J. Complexity* **54** (2019), pages 101394, 17. DOI: 10.1016/j.jco.2019.01.001.

[Pro92] Yurii V. Prokhorov. 'On polynomials in normally distributed random variables'. In: *Teor. Veroyatnost. i Primenen.* **37.4** (1992), pages 747–750. DOI: 10.1137/1137123.

[Pro93] Yurii V. Prokhorov. 'On polynomials in random variables with Gamma-distribution'. In: *Teor. Veroyatnost. i Primenen.* **38.1** (1993), pages 198–202. DOI: 10.1137/1138015.

[RS14] Maxim Raginsky and Igal Sason. *Concentration of measure inequalities in information theory, communications, and coding.* Now Publishers Inc., 2014, pages 1–260. ISBN: 978-1-60198-906-2.

[RR09] Gesine Reinert and Adrian Röllin. 'Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition'. In: *Ann. Probab.* **37.6** (2009), pages 2150–2173. DOI: 10.1214/09-AOP467.

[Rio01] Emmanuel Rio. 'Inégalités de concentration pour les processus empiriques de classes de parties'. In: *Probab. Theory Related Fields* **119.2** (2001), pages 163–175. DOI: 10.1007/PL00008756.

[Rio02] Emmanuel Rio. 'Une inégalité de Bennett pour les maxima de processus empiriques'. In: *Ann. Inst. H. Poincaré Probab. Statist.* **38.6** (2002), pages 1053–1057. DOI: 10.1016/S0246-0203(02)01122-6.

[Roy07] Gilles Royer. *An initiation to logarithmic Sobolev inequalities.* Volume 14. SMF/AMS Texts and Monographs. American Mathematical Society, Providence, RI, 2007, pages viii+119. ISBN: 978-0-8218-4401-4.

[Ruc88] Andrzej Ruciński. 'When are small subgraphs of a random graph normally distributed?' In: *Probab. Theory Related Fields* **78.1** (1988), pages 1–10. DOI: 10.1007/BF00718031.

[RV13] Mark Rudelson and Roman Vershynin. 'Hanson–Wright inequality and sub-Gaussian concentration'. In: *Electron. Commun. Probab.* **18** (2013), no. 82, 9. DOI: 10.1214/ECP.v18-2865.

[SS18] Holger Sambale and Arthur Sinulis. 'Logarithmic Sobolev inequalities for finite spin systems and applications'. In: *arXiv preprint* (2018). accepted for publication in *Bernoulli*. arXiv: 1807.07765.

[SS19] Holger Sambale and Arthur Sinulis. 'Modified log-Sobolev inequalities and two-level concentration'. In: *arXiv preprint* (2019). arXiv: 1905.06137.

[Sam17] Paul-Marie Samson. 'Transport-entropy inequalities on locally acting groups of permutations'. In: *Electron. J. Probab.* **22** (2017), no. 62, 33. DOI: 10.1214/17-EJP54.

[SS11] Warren Schudy and Maxim Sviridenko. 'Bernstein-like concentration and moment inequalities for polynomials of independent random variables: multilinear case'. In: *arXiv preprint* (2011). arXiv: 1109.5193.

[SS12] Warren Schudy and Maxim Sviridenko. 'Concentration and moment inequalities for polynomials of independent random variables'. In: *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, 2012, pages 437–446. DOI: 10.1137/1.9781611973099.37.

[Ste86] J. Michael Steele. 'An Efron–Stein inequality for nonsymmetric statistics'. In: *Ann. Statist.* **14.2** (1986), pages 753–758. DOI: 10.1214/aos/1176349952.

[Str86] David Strauss. 'On a general class of models for interaction'. In: *SIAM Rev.* **28.4** (1986), pages 513–527. DOI: 10.1137/1028156.

[ST74] Vladimir N. Sudakov and Boris S. Tsirel'son. 'Extremal properties of half-spaces for spherically invariant measures'. In: *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **41** (1974), pages 14–24. DOI: 10.1007/BF01086099.

[Tal88] Michel Talagrand. 'An isoperimetric theorem on the cube and the Kintchine-Kahane inequalities'. In: *Proc. Amer. Math. Soc.* **104.3** (1988), pages 905–909. DOI: 10.2307/2046814.

[Tal91a] Michel Talagrand. 'A new isoperimetric inequality and the concentration of measure phenomenon'. In: *Geometric Aspects of Functional Analysis*. Volume 1469. Lecture Notes in Mathematics. Springer, Berlin, 1991, pages 94–124. DOI: 10.1007/BFb0089217.

[Tal91b] Michel Talagrand. 'A new isoperimetric inequality for product measure and the tails of sums of independent random variables'. In: *Geom. Funct. Anal.* **1.2** (1991), pages 211–223. DOI: 10.1007/BF01896379.

[Tal95] Michel Talagrand. 'Concentration of measure and isoperimetric inequalities in product spaces'. In: *Inst. Hautes Études Sci. Publ. Math.* **81** (1995), pages 73–205. DOI: 10.1007/BF02699376.

[Tal96a] Michel Talagrand. 'A new look at independence'. In: *Ann. Probab.* **24.1** (1996), pages 1–34. DOI: 10.1214/aop/1042644705.

[Tal96b] Michel Talagrand. 'New concentration inequalities in product spaces'. In: *Invent. Math.* **126.3** (1996), pages 505–563. DOI: 10.1007/s002220050108.

[Tro12] Joel A. Tropp. 'User-friendly tail bounds for sums of random matrices'. In: *Found. Comput. Math.* **12.4** (2012), pages 389–434. DOI: 10.1007/s10208-011-9099-z.

[Tro15] Joel A. Tropp. 'An introduction to matrix concentration inequalities'. In: *arXiv preprint* (2015). arXiv: 1501.01571.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009, pages xii+214. ISBN: 978-0-387-79051-0.

[Ver12] Roman Vershynin. 'Introduction to the non-asymptotic analysis of random matrices'. In: *Compressed sensing*. Cambridge University Press, Cambridge, 2012, pages 210–268. DOI: 10.1017/CBO9780511794308.006.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Volume 47. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018, pages xiv+284. ISBN: 978-1-108-41519-4.

[Vil09] Cédric Villani. *Optimal transport. Old and new*. Volume 338. Grundlehren der Mathematischen Wissenschaften. Springer-Verlag, Berlin, 2009, pages xxii+973. ISBN: 978-3-540-71049-3.

[Vu02] Van H. Vu. 'Concentration of non-Lipschitz functions and applications'. In: *Random Structures Algorithms* **20.3** (2002), pages 262–316. DOI: `10.1002/rsa.10032`.

[VW15] Van H. Vu and Ke Wang. 'Random weighted projections, random quadratic forms and random eigenvectors'. In: *Random Structures Algorithms* **47.4** (2015), pages 792–821. DOI: `10.1002/rsa.20561`.

[Wol13] Paweł Wolff. 'On some Gaussian concentration inequality for non-Lipschitz functions'. In: *High Dimensional Probability VI*. Volume 66. Progr. Probab. Birkhäuser, Basel, 2013, pages 103–110. DOI: `10.1007/978-3-0348-0490-5_8`.

[Wri73] Farroll T. Wright. 'A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric'. In: *Ann. Probability* **1.6** (1973), pages 1068–1070. DOI: `10.1214/aop/1176996815`.

[Yan+19] Zhuoran Yang, Lin F. Yang, Ethan X. Fang, Tuo Zhao, Zhaoran Wang, and Matey Neykov. 'Misspecified nonconvex statistical optimization for sparse phase retrieval'. In: *Math. Program.* **176.1-2, Ser. B** (2019), pages 545–571. DOI: `10.1007/s10107-019-01364-5`.