Second Order Concentration for Functions of Independent Random Variables

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Mathematik
der Universität Bielefeld

vorgelegt von
Holger Sambale

30.11.2015

betreut durch Herrn Prof. Dr. Friedrich Götze

Contents

1 Introduction

The starting point of this thesis is the following situation: consider n independent random variables $X_1, \dots, X_n$ and a bounded measurable function $f\colon \mathbb{R}^n \to \mathbb{R}$. Then, we want to deduce concentration inequalities of second order for the statistic $f(X_1, \dots, X_n)$.

In doing so, we lean on the work by S. G. Bobkov, G. P. Chistyakov and F. Götze [B-C-G] which is about second order concentration inequalities for the n-dimensional unit sphere equipped with the uniform distribution. In fact, in our situation as presented above the special case of all random variables having Bernoulli distribution $\mu = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ provides an analogue of the results in [B-C-G] for the discrete hypercube. Actually, it was this situation which served as a first motivation for this work.

We will give a more comprehensive overview about the background of this work in Section 2. In particular, we will discuss several related results at that point.

The meaning of "second order concentration" is developed similarly to [B-C-G]. Instead of using derivatives, we will introduce suitable difference operators. Moreover, we will work with the Hoeffding decomposition of $f(X_1, \dots, X_n)$. All this will be discussed in Section 3. Second order concentration then means that the inequalities we derive will make use of and depend on difference operators of second order. Furthermore, they will usually require functions whose Hoeffding decompositions only begin with terms of second order.

In the situation on the unit sphere as discussed in [B-C-G], concentration inequalities are derived by making use of the fact that the uniform distribution on the sphere satisfies a logarithmic Sobolev inequality. This is mirrored in our work by working with modified logarithmic Sobolev inequalities which make use of the difference operators introduced in Section 3. This will be done in Section 4.

We now formulate our central results:

Theorem 1.1. *Let $\mu_1, \dots, \mu_n$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and denote by $\mu = \otimes_{i=1}^n \mu_i$ their product measure. Moreover, let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a measurable function which is bounded on the support of μ so that its Hoeffding decomposition with respect to μ is given by*

$$f = \sum_{k=d}^{n} f_k$$

for some $d \geq 2$. Denote by D and ∇ the difference operators as introduced in Example 3.2. Assume that the condition

$$|\nabla|\nabla f|| \leq 1$$

is satisfied on the support of μ and that we have

$$\int \|f^{\hat{''}}\|_{\mathrm{HS}}^2 d\mu \leq b^2$$

for some $b \geq 0$. Here, $f^{\hat{''}}$ is the "dediagonalized" Hessian of f with respect to the difference operator D, and $\|f^{\hat{''}}\|_{\mathrm{HS}}$ denotes its Hilbert Schmidt norm.

Then, we have

$$\int \exp\left(\frac{1}{2(3+b^2/(d-1))}|f|\right) d\mu \leq 2.$$

If all the measures μ_i in Theorem 1.1 are Bernoulli measures, we can sharpen the bound given above a little. More precisely, we have:

Theorem 1.2. *In the situation of Theorem 1.1, let all the μ_i be of the form $\mu_i = p_i\delta_{+1} + (1-p_i)\delta_{-1}$ with δ_x denoting the Dirac measure at $x \in \mathbb{R}$ and $p_i \in (0,1)$ for all i. Then, with the conditions given in Theorem 1.1, we have*

$$\int \exp\left(\frac{1}{3+2b^2/(d-1)}|f|\right) d\mu \leq 2.$$

The choice of $\{\pm1\}^n$ as the underlying space in Theorem 1.2 is motivated by the discrete hypercube. In general, we can take any two-point measures on $(\mathbb{R}, \mathbb{B})$ possibly even varying from one component to the other.

Using Chebychev's inequality, Theorems 1.1 and 1.2 for instance imply the estimate

$$\mu(|f| \geq t) \leq 2e^{-ct}$$

for all $t > 0$ and some constant $c = c(b^2, d)$. The value of the latter constant as given by the bounds in Theorems 1.1 and 1.2 is not optimal, but optimizing it seems hard. It is possible to obtain a slightly better but still non-optimal constant from the proof of Theorems 1.1 and 1.2.

It is possible to generalize Theorem 1.1 by considering probability measures μ_i on arbitrary measurable spaces $(X_i, \mathcal{X}_i)$ for all $i = 1, \ldots, n$. The theorem and its proof can easily be adapted to this situation. By contrast, it is not possible to remove the boundedness condition on f by our methods. Indeed, in the appendix we will prove that requiring $|\nabla|\nabla f|| \leq 1$ as in Theorem 1.1 already implies that f must be bounded. Still, we might try to obtain inequalities for unbounded functions as a sort of "limit" from those for bounded functions. However, we cannot expect to arrive at any useful results (consider the case of $\tau \to \infty$ in the following Remark 1.3).

In some applications, it is convenient not to consider "normalized" functions f in the sense of requiring $|\nabla|\nabla f|| \leq 1$ on the support of μ but to allow a general upper bound $\tau \geq 0$. Theorems 1.1 and 1.2 are easily adapted to this situation. For convenience, we state this explicitly:

Remark 1.3. *In the situation of Theorem 1.1 or Theorem 1.2, replace the condition* $|\nabla|\nabla f|| \leq 1$ *by* $|\nabla|\nabla f|| \leq \tau$ *on the support of* μ *for some* $\tau \geq 0$*. Then, we can rewrite our results as*

$$\int \exp\left(\frac{1}{2(3\tau + \tau^{-1}b^2/(d-1))}|f|\right) d\mu \leq 2$$

or, in the Bernoulli case,

$$\int \exp\left(\frac{1}{3\tau + 2\tau^{-1}b^2/(d-1)}|f|\right) d\mu \leq 2.$$

In particular, if we have $b = \tau$ *and* $d = 2$*, we get*

$$\int e^{|f|/(8\tau)} d\mu \leq 2 \qquad \text{or} \qquad \int e^{|f|/(5\tau)} d\mu \leq 2,$$

respectively.

The proof of Theorem 1.1 once again leans on [B-C-G] and will be given in Sections 5 and 6. This bisection is a consequence of the proof mainly consisting of two steps. First, we will derive exponential inequalities involving the difference operator ∇ by making use of modified logarithmic Sobolev inequalities. After that, we will relate ∇ to second order difference operators D_{ij} in form of a "Hessian". Here, one tool is an appropriately defined Laplacian. Having established all this, the proof of Theorem 1.1 and 1.2 is easily obtained by combining both streams.

So far, we have considered functions whose Hoeffding decomposition with respect to the underlying measures starts with (at least) second order terms. In Section 7, we will formulate a slightly generalized version of Theorem 1.1 in which we also allow Hoeffding terms of first order if certain extra conditions are satisfied. This will however entail some inconveniences.

Finally, we present some applications of Theorem 1.1. In Section 8, we begin with the easiest case, namely functions in independent symmetric Bernoulli random variables, in order to check how our results work in a simple situation. In particular, we demonstrate how the condition $|\nabla|\nabla f|| \leq 1$ can be transferred into a condition on the Hilbert-Schmidt norm of $f^{\hat{''}}$.

In Section 9, we will then generalize these results to multilinear polynomials in independent random variables (given some restrictions like boundedness). This is partly inspired by E. Mossel, R. O'Donnell and K. Oleszkiewicz [M-O-O], even if our point of view is a different one and in particular our work does not aim at an invariance principle as in the latter article. We will see that we can transfer our results from the symmetric Bernoulli case to the more general situation, but the fact that we do not have $X_i^2 \equiv 1$ anymore will cause some additional work, which in particular results in a further condition.

In Section 10, we continue with second order concentration of empirical distribution functions. Here, we lean on [B-G3] and transport some of their results to the situation of Theorem 1.1. This includes the behavior of the Kolmogorov distance of the empirical distribution function and the mean empirical distribution function.

A partial application of the results from Section 10 is given in Section 11. We will consider empirical distribution functions which are based on a set of independent symmetric Bernoulli variables. This will enable us to use the methods from Section 8. We will obtain some simple results about second order concentration of such distributions as, for instance, typical situations in which we can well apply our results.

In Section 12, we finally present another type of application which is more advanced than some of the previous examples. We consider Erdős-Rényi random graphs $G(n, p)$, i. e. roughly speaking graphs that consist of n vertices such that there is an edge between any two of them with probability p and the edges are chosen independently. In this setting, an interesting problem is to count the number of subgraphs which are contained in $G(n, p)$, and the subgraphs we will focus on are triangles. In particular, we compare our results to those stated in Adamczak and Wolff [A-W]. In fact, this part of the thesis is primarily inspired by the latter article though there is much other literature related to this topic.

Acknowledgements. I am very grateful to my advisor Professor Dr. Friedrich Götze for suggesting such an interesting and challenging project. This thesis could not have been written without his guidance and support. Moreover, I would like to thank Dr. Holger Kösters for many fruitful discussions. Finally, I want to thank Professor Dr. Sergey Bobkov for some inspiring ideas which helped to make this thesis round.

2 Related Work

This work is related to two big fields in probability theory. One of them is the concentration of measure phenomenon with a special focus on logarithmic Sobolev inequalities, the other one is the limit behavior of U-statistics.

The concentration of measure phenomenon dates back to the 1970s, where it was particularly highlighted in the work of V. N. Sudakov and V. D. Milman. A typical example is the Gaussian concentration property, which states that for any Borel set $A \subset \mathbb{R}^n$ with standard Gaussian measure $\gamma(A) \geq 1/2$, we have

$$\gamma(A_r) \leq 1 - e^{-r^2/2}$$

for any $r \geq 0$ and $A_r := \{x \in \mathbb{R}^n : d(x, A) < r\}$, where d is the Euclidean distance. This can be reformulated in terms of Lipschitz functions. In detail, we get that for any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ with Lipschitz constant at most 1, we have

$$\gamma(f - \int f d\gamma \geq r) \leq e^{-r^2/2}$$

for any $r \geq 0$.

These results imply a number of further questions as for instance which other measures may satisfy similar concentration properties. One powerful tool for studying problems of such type are logarithmic Sobolev inequalities, which were introduced at about the same time as the concentration of measure phenomenon. In particular, logarithmic Sobolev inequalities for the symmetric Bernoulli measure as well as for Gaussian measures were introduced and proved by L. Gross [G] in 1975.

Since then, much research has been done in this area, and a variety of methods and applications has been developed. We do not give a detailed account but only focus on those which are particularly important for our own work. For a comprehensive survey which summarizes the central results up to the end of the 1990s see the monographs by M. Ledoux [L2], [L3].

One of the basic tools we use are techniques which make it possible to deduce concentration inequalities from logarithmic Sobolev inequalities with the help of Laplace transforms. This was first seen by I. Herbst and developed further by S. Aida, T. Masuda and I. Shikegawa [A-M-S].

The discussion of the concentration of measure phenomenon for the case of product measures was particularly put forward by M. Talagrand in the 1990s, resulting in papers like [T1] and [T2]. It was subsequently taken up by others like S. Bobkov and M. Ledoux. For our own work, the results of S. G. Bobkov and F. Götze [B-G1] are of particular importance, since our way of introducing modified logarithmic Sobolev inequalities and some of the exponential concentration results are based on this article.

The second block of results which are used in this work is about U-statistics, that is, statistics of the form

$$U_n(h) = \frac{1}{\binom{n}{m}} \sum_{i_1 < \ldots < i_m} h(X_{i_1}, \ldots, X_{i_m})$$

for a sequence of i.i.d. random variables $(X_i)_{i \in \mathbb{N}}$, a measurable (kernel) function h on $\mathbb{R}^m$ and natural numbers n, m such that $n \geq m$. Such statistics allow a decomposition which was introduced by W. Hoeffding [H] in 1948 and which is called the Hoeffding decomposition. This decomposition is orthogonal if $h(X_1, \ldots, X_m)$ is square-integrable.

From the late 1940s on, much research has been done in this area, and the Hoeffding decomposition is now a classical tool for analyzing the distributional properties of U-statistics. An overview about the main results is partly given in the monograph by V. de la Peña and E. Giné [D-G].

In the context of this thesis, we especially refer to the many inequalities which study the tail behavior of U-statistics. This goes back to Hoeffding's inequalities as stated in Theorem 4.1.8 in [D-G], which in particular yields that for $U_n(h)$ as defined above, we have

$$\mathbb{P}(U_n(h) > t) \leq \exp\left(-\frac{[n/m]t^2}{2M^2}\right)$$

if the function $h\colon \mathbb{R}^m \to \mathbb{R}$ is bounded by some universal constant M and satisfies $\mathbb{E}h(X_1, \dots, X_m) = 0$. Note that this inequality can be regarded a first order analogue of our own work.

Later authors have studied exponential inequalities for U-statistics which are completely degenerate (or canonical), which means their Hoeffding decomposition consists of a single term only. An overview can once again be found in de la Peña and Giné [D-G], Chapter 4.1.3. In particular, we mention M. A. Arcones and E. Giné [A-G] and M. A. Arcones again [A] as well as the results by P. Major [M] (e. g. Theorem 8.3 in [M]). In [A-G] and [M], tail inequalities for completely degenerate U-statistics are given such that the estimates only depend on the order m of the kernel h, its second moment σ^2 and some bound M on h.

In [A], a bounded, centered but not necessarily degenerated kernel h is considered, for which improved tail estimates are deduced by replacing the variance of h by the variance of the first order Hoeffding term of $U_n(h)$. The basic idea of the proof is separating the first order Hoeffding term from the remaining ones, which resembles one of the key ideas of our own work. However, the goals and techniques used in [A] are quite different from our own ones.

In particular, a central difference is that while the results in [A] require statistics with Hoeffding decompositions which stop at some order m (which should be independent of n to get useful results), the statistics we consider in Theorem 1.1 can have terms from order 2 up to n. We will continue this discussion in Sections 8 and 9.

To conclude this section, there are at least two results which we highlight explicitly because there is a closer relation to our work.

First, a first order analogue of our work which is even more immediate than Hoeffding's tail estimates can be found in [B-G2], for instance. In principle, similar results were already known at an earlier point as e. g. in [L1], but in [B-G2] a systematic account is provided on which we lean in our own work. For completeness, we now give a slightly reformulated version of Proposition 2.1 from there:

Proposition 2.1. *Let $\mu_1, \dots, \mu_n$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and denote by $\mu = \otimes_{i=1}^n \mu_i$ their product measure. Moreover, let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a measurable function which is bounded on the support of μ. Denote by ∇ the difference operator as introduced in Example 3.2.2. Assume that the condition $|\nabla f| \leq 1$ is satisfied on the support of μ.*

Then, for any $t \geq 0$ we have

$$\mu\left(|f - \int f d\mu| \geq t\right) \leq 2e^{-t^2/4}.$$

As compared to the original result in [B-G2], we have added the assumption that f is bounded, which does not change anything since as we have shown in the appendix, the condition $|\nabla f| \leq 1$ implies the boundedness of f anyway. The slight differences on the right hand side are due to the fact that we have restricted ourselves to real-valued functions (while in [B-G2] complex-valued function are considered as well).

It is not yet necessary to introduce the notion of Hoeffding decomposition or a second type of difference operator in Proposition 2.1. On the other hand, subtracting the expectation $\int f d\mu$ means removing the Hoeffding term of order zero (cf. Theorem 3.4), so that we can recognize the basic structure of Theorem 1.1 in Proposition 2.1 as well.

As for applications of Proposition 2.1, we refer to [L1], where such results (even if formulated in a slightly different way) are used to study concentration in product spaces which are equipped with Hamming metrics (i. e. $d(x, y) := \operatorname{card}\{k = 1, \ldots, n \colon x_k \neq y_k\}$). In the same paper, they are also applied in the context of "penalties" (which were introduced in [T1] and can be regarded as generalizations of the Hamming metric).

Moreover, in [B-G2], a generalized version of Proposition 2.1 (i. e. for complex-valued functions) is used as a tool for deducing concentration inequalities for randomized sums. Here the complex-valued case is needed because the concentration inequalities are applied to characteristic functions.

A second important collection of results which are related to our work are those by R. Adamczak and P. Wolff [A-W]. For our work, Theorems 1.2 and 1.4 from [A-W] are particularly interesting. In Theorem 1.2, the underlying measure must fulfill a certain Sobolev-type inequality, i. e.

$$\|f(X) - \mathbb{E}f(X)\|_p \leq L\sqrt{p}\||\nabla f(X)|\|_p \qquad (*)$$

for any $p \geq 2$, any smooth integrable function $f \colon \mathbb{R}^n \to \mathbb{R}$ and some constant L which does not depend on p and f. Here, X is a random vector with the distribution in question, and $|\cdot|$ is the standard Euclidean norm.

In particular, Theorem 1.2 in [A-W] then states that for any random vector X in $\mathbb{R}^n$ satisfying $(*)$ and any function $f \colon \mathbb{R}^n \to \mathbb{R}$ in $\mathcal{C}^k$ such that $D^k f(x)$ is uniformly bounded on $\mathbb{R}^n$, we have

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\left(-\frac{1}{C_k}\eta_f(t)\right)$$

for any $t > 0$, where C_k is some constant, and

$$\eta_f(t) = \min\left(\min_{\mathcal{J}\in P_k}\left(\frac{t}{L^k \sup_{x\in\mathbb{R}^n}\|D^k f(x)\|_{\mathcal{J}}}\right)^{\frac{2}{\#\mathcal{J}}}, \min_{1\leq d\leq k-1}\min_{\mathcal{J}\in P_d}\left(\frac{t}{L^d\|\mathbb{E}D^d f(x)\|_{\mathcal{J}}}\right)^{\frac{2}{\#\mathcal{J}}}\right).$$

Here, the derivatives $D^k f(x)$ are regarded as multi-indexed matrices, L is the constant from $(*)$, P_d denotes the group of the partitions of $\{1,\dots,d\}$, and $\|\cdot\|_{\mathcal{J}}$, $\mathcal{J} \in P_d$, is some family of tensor-product matrix norms.

However, requiring $(*)$ excludes the case of discrete measures we are especially interested in. In this context, Theorem 1.4 from [A-W] seems a more immediate comparison to our work since here, products of measures which have subgaussian tail decay are considered. In particular, Theorem 1.4 from [A-W] states the following: Let $X = (X_1,\dots,X_n)$ be a random vector with independent components such that for all $i \leq n$ we have

$$\|X_i\|_{\psi_2} := \inf\{t > 0 \colon \mathbb{E}\exp\left(\frac{X_i^2}{t^2}\right) \leq 2\} \leq L$$

for some $L \geq 0$. Then, for every polynomial $f\colon \mathbb{R}^n \to \mathbb{R}$ of degree k and any $t > 0$, we have

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\left(-\frac{1}{C_k}\eta_f(t)\right),$$

where

$$\eta_f(t) = \min_{1\leq d\leq k}\min_{\mathcal{J}\in P_d}\left(\frac{t}{L^d\|\mathbb{E}D^d f(x)\|_{\mathcal{J}}}\right)^{\frac{2}{\#\mathcal{J}}}.$$

However, this theorem only allows polynomial functions, while our results hold for arbitrary but bounded functions. Moreover, the notion of Hoeffding decomposition which is a central aspect in our work does not appear in [A-W]. Altogether, we place a stronger emphasis on the discrete situation than [A-W] do.

Finally, as we already did in Section 1, we should once again mention the work of S. G. Bobkov, G. P. Chistyakov and F. Götze [B-C-G], which was the prime source of inspiration for this thesis. Theorem 1.1 and its proof are in principle an adaption of the results for functions on the unit sphere as discussed in [B-C-G]. In this context, a central task of the present work was finding the definitions which are best suitable for working in a similar way as [B-C-G].

3 Difference Operators

Our first step in the course of deducing Theorem 1.1 is introducing several types of difference operators. They serve as an analogue of the spherical derivatives in [B-C-G].

The most important type of difference operators used in our work are operators Γ on the space of the bounded measurable real-valued functions on $(\mathbb{R}^n, \mathbb{B}^n)$ such that the following two conditions hold:

Conditions 3.1. *(i) For any bounded measurable function* $f\colon \mathbb{R}^n \to \mathbb{R}$, $\Gamma f = (\Gamma_1 f, \dots \Gamma_n f)\colon \mathbb{R}^n \to \mathbb{R}^n$ *is a measurable function with values in* $\mathbb{R}^n$. *We often call* Γ *a* gradient operator *or simply* gradient.

(ii) For all $i = 1, \dots, n$, *all* $a > 0$, $b \in \mathbb{R}$ *and any bounded measurable real-valued function* f, *we have* $|\Gamma_i(af + b)| = a|\Gamma_i f|$.

These conditions are basically due to Bobkov and Götze [B-G1]. In particular, we do not suppose Γ to satisfy any sort of "Leibniz rule". Note that we require the functions f to be bounded just in order to avoid too many extra assumptions.

It is easily possible to rewrite Conditions 3.1 such that they hold for arbitrary measurable spaces $(X, \mathcal{X})$ (corresponding to the case $n = 1$) or finite products of measurable spaces $(X_i, \mathcal{X}_i)$. In Sections 4 and 5 we will mostly work in such a general situation.

We now give three examples of difference operators which we will need in our further discussion. Each of them is based on the assumption that $(\mathbb{R}^n, \mathbb{B}^n)$ is endorsed with some product probability measure.

Example 3.2. *1. Let* μ_i, $i = 1, \dots, n$, *be probability measures on* $(\mathbb{R}, \mathbb{B})$, $\mu = \otimes_{i=1}^n \mu_i$ *their product measure and* $f\colon \mathbb{R}^n \to \mathbb{R}$ *a* μ*-integrable function. Then, we define*

$$D_i f(x) := f(x) - \int_{\mathbb{R}} f(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n)\mu_i(dy_i), \tag{1}$$

i. e. we subtract the integral with respect to the i-th component. We then define the respective gradient operator by $Df := (D_1 f, \dots, D_n f)$. *We will sometimes call this type of difference operator "Hoeffding difference".*

Based on the first order differences $D_i f$, *we define higher order differences by iteration, i. e. for instance* $D_{ij} f := D_i(D_j f)$ *for* $1 \le i, j \le n$.

2. In the situation of Part 1 but for f *being an* $L^2(\mu)$*-function, we set*

$$\nabla_i f(x) := \Big(\frac{1}{2}\int_{\mathbb{R}} (f(x) - f(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n))^2 \mu_i(dy_i)\Big)^{1/2} \tag{2}$$

and as in Part 1 $\nabla f := (\nabla_1 f, \dots, \nabla_n f)$.

3. In the same situation as in Part 2, we define

$$\nabla_i^+ f(x) := \Big(\frac{1}{2}\int_{\mathbb{R}} (f(x) - f(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n))_+^2 \mu_i(dy_i)\Big)^{1/2}, \tag{3}$$

where for any real-valued function g *we set* $g_+ := \max(g, 0)$ *for its positive part.*

In all cases, Conditions 3.1 are clearly satisfied. Note again that we can easily adapt the definitions of the three operators given above to the situation of arbitrary measurable (product) spaces. For instance, such a general definition of the operator ∇ is given in [B-G2].

The choice of the factor $1/2$ in the definitions of ∇ and ∇^+ is arbitrary in principle, and we could choose any other positive real number instead. However, if we choose $1/2$, we can relate all three types of difference operators from Example 3.2 if the underlying measure is the uniform distribution on a two-point space:

Remark 3.3. *1. Consider the case where $\mu_i = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ for all $i = 1, \ldots, n$, i. e. the uniform distribution on $\{\pm 1\}^n$. Then, we have*

$$D_i f(x) = \frac{1}{2}(f(x) - f(\sigma_i x))$$

with $\sigma_i x := (x_1, \ldots, -x_i, \ldots, x_n)$ for any $x \in \{\pm 1\}^n$. Moreover, we have the relations $\nabla_i f = |D_i f|$ as well as $\nabla_i^+ f = (D_i f)_+$.

2. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X_1, \ldots, X_n$ independent random variables on it with distributions μ_i, $i = 1, \ldots, n$. Then, we can rewrite (1) as

$$D_i f(x) = f(x) - \mathbb{E}_i f(x)$$

or (in short) $D_i = Id - \mathbb{E}_i$. Here, Id denotes the identity and $\mathbb{E}_i$ taking the expectation with respect to X_i. (Strictly spoken, we interpret $\mathbb{E}_i f(x)$ as replacing the entry x_i in the i-th component of x by the random variable X_i and then taking the expectation.)

3. Similarly to the description of D_i given above, we can also rewrite ∇_i in terms of expectations of certain random variables. For that, consider the situation as described above and add a set of independent copies $\bar{X}_1, \ldots, \bar{X}_n$ of the random variables $X_1, \ldots, X_n$. For a function $f(X_1, \ldots, X_n)$ set $T_i f := f(X_1, \ldots, X_{i-1}, \bar{X}_i, X_{i+1}, \ldots, X_n)$, i. e. the random variable X_i is replaced by its independent copy $\bar{X}_i$. Then, we have

$$\nabla_i f(x) = \Big(\frac{1}{2}\bar{\mathbb{E}}_i (f(x) - T_i f(x))^2\Big)^{1/2}.$$

Here, $\bar{\mathbb{E}}_i$ means taking the expectation with respect to $\bar{X}_i$.

4. In the same vein as in Part 3, we can write

$$\nabla_i^+ f(x) = \Big(\frac{1}{2}\bar{\mathbb{E}}_i (f(x) - T_i f(x))_+^2\Big)^{1/2}$$

as well as

$$D_i f(x) = \bar{\mathbb{E}}_i (f(x) - T_i f(x)).$$

We will not worry much about making notation as precise as possible in this work. For instance, we will sometimes write $T_i f(X_1, \ldots, X_n)$ though being aware of the i-th component having actually been replaced by an independent copy. Partially we may even drop the arguments completely (as we did above). Also, we will switch from integrals with respect to the distributions μ_i to expected values with random variables involved whenever this seems convenient.

Furthermore, in the sequel we will occasionally make use of the fact that the composition of T_i and a function $g\colon \mathbb{R} \to \mathbb{R}$ is commutative, i. e. we have identities of the form $T_i f^2 = (T_i f)^2$ (with $g(x) := x^2$ in this case) and similar relations.

For our work, we also need an analogue of the Laplacian for difference operators. Here we use the operator D from Example 3.2.1 since it can easily be iterated. Note that according to Example 3.2.1, we have $D_{ii} = D_i$ for all i. Therefore, we do not adapt the definition of the "ordinary" Laplacian but consider second order differences in two *different* directions instead. That is, we set

$$\Delta := \sum_{i \neq j} D_{ij}. \tag{4}$$

Calling (4) a Laplacian is justified for several reasons. First, if we compare our work to [B-C-G] once again we see that (4) plays the same role in our arguments as the spherical Laplacian does in [B-C-G].

Moreover, it is well-known that the usual Laplacian on $\mathbb{R}^n$ is invariant under rotations. In discrete situations, we can regard invariance under permutations as an analogue of this property. Indeed, if we assume $\mu_i \equiv \mu_1$ for all i in Example 3.2, i. e. the u.i.v. case if we consider random variables, the Laplacian (4) indeed satisfies the relation

$$\Delta f(x) = \Delta f(\pi(x)),$$

where f is any μ-integrable function on $\mathbb{R}^n$ and π is any permutation of $\{1, 2, \ldots, n\}$. As usual, here we set $f(\pi(x)) = f(x_{\pi^{-1}(1)}, \ldots, x_{\pi^{-1}(n)})$.

On the other hand, some caution is necessary, for it seems (4) is not a Laplacian in the stochastic sense. That is, we have not been able to find a homogeneous Markov chain such that (4) is the corresponding Laplacian, and it seems this is not possible in fact. Nevertheless, for the reasons sketched above we believe that calling (4) a Laplacian makes sense in our context.

Next we introduce the notion of *Hoeffding decomposition*. This can be done as follows:

Theorem 3.4. *Let $X_1, \dots, X_n$ be independent random variables on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function such that $f(X_1, \dots, X_n)$ is integrable. Then, there is a unique decomposition*

$$\begin{aligned} f(X_1, \dots, X_n) &= \mathbb{E}f + \sum_{i=1}^{n} h_i(X_i) + \sum_{i<j} h_{ij}(X_i, X_j) + \dots \\ &= f_0 + f_1 + f_2 + \dots + f_n \end{aligned}$$

such that $\mathbb{E}_{i_j} h_{i_1 \dots i_k}(X_{i_1}, \dots, X_{i_k}) = 0$ for all $k = 1, \dots, n$, $1 \le i_1 < \dots < i_k \le n$ and $j \in \{1, \dots, k\}$. Here, as above, $\mathbb{E}_i$ means integration with respect to X_i. This decomposition is called the Hoeffding decomposition, *and the sum f_d is called the* Hoeffding term of degree d *or simply* d-th Hoeffding term *of f.*

Proof. As this is a well-known result, we only give a brief sketch of the proof. Let D_i, $i = 1, \dots n$, be the difference operators we defined in (1). Then, $\{\mathbb{E}_i, i = 1, \dots, n\} \cup \{D_i, i = 1, \dots, n\}$ is a family of commutative operators with respect to composition, and we clearly have $\mathbb{E}_i + D_i = Id$. Hence, for a function f with properties as in the theorem, we get (notating composition as multiplication)

$$f(X_1, \dots, X_n) = \prod_{i=1}^{n} (\mathbb{E}_i + D_i) f(X_1, \dots, X_n).$$

Expanding the product then gives the desired decomposition with

$$h_{i_1 \dots i_k}(X_{i_1}, \dots, X_{i_k}) = \Big(\prod_{j \notin \{i_1, \dots i_k\}} \mathbb{E}_j \prod_{l \in \{i_1, \dots i_k\}} D_l \Big) f(X_1, \dots, X_n).$$

The uniqueness of this decomposition is easily seen. □

As a consequence, if $f(X_1, \dots, X_n) \in L^2$, then the f_k are pairwise uncorrelated, i. e. orthogonal in L^2. If we now relate the Hoeffding decomposition to the Laplacian (4), we observe that the Hoeffding terms are eigenfunctions of the Laplace operator. Thus, as an analogue of the situation on the sphere as discussed in [B-C-G], in our case we also have an orthogonal decomposition on which Δ operates diagonally.

Theorem 3.5. *Let $X_1, \dots, X_n$ be independent random variables on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function such that $f(X_1, \dots, X_n)$ is integrable. Moreover, let $f = \sum_{d=0}^{n} f_d$ be the Hoeffding decomposition of f. Then, we have*

$$\Delta f_d = (d)_2 f_d.$$

Here, Δ is the Laplacian as introduced in (4), and we write $(d)_2 = d(d-1)$. Thus, the d-th Hoeffding term is an eigenfunction of Δ with eigenvalue $(d)_2$.

Proof. Let $f_d(X_1, \dots, X_n) = \sum_{i_1 < \dots < i_d} h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})$. Fix $i_1 < \dots < i_d$. Then, we get

$$\mathbb{E}_i h_{i_1 \dots i_d}(X_{i_1} \dots, X_{i_d}) = \begin{cases} 0, & i \in \{i_1, \dots, i_d\}, \\ h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d}), & i \notin \{i_1, \dots, i_d\}, \end{cases}$$

due to the properties of the Hoeffding decomposition as stated in Theorem 3.4. Keeping in mind that $D_i = Id - \mathbb{E}_i$ (cf. Remark 3.3.2), we therefore immediately obtain

$$D_i f_d(X_1, \ldots, X_n) = \sum_{\substack{i_1 < \ldots < i_d \\ i \in \{i_1, \ldots, i_d\}}} h_{i_1 \ldots i_d}(X_{i_1}, \ldots, X_{i_d}) \tag{5}$$

and consequently

$$D_{ij} f_d(X_1, \ldots, X_n) = \sum_{\substack{i_1 < \ldots < i_d \\ i,j \in \{i_1, \ldots, i_d\}}} h_{i_1 \ldots i_d}(X_{i_1}, \ldots, X_{i_d}). \tag{6}$$

So it remains to check how often each term $h_{i_1 \ldots i_d}(X_{i_1}, \ldots, X_{i_d})$ appears in $\Delta f_d = \sum_{i \neq j} D_{ij} f_d$. As we just saw, each pair $i \neq j$ such that $i, j \in \{i_1, \ldots, i_d\}$ replicates the summand $h_{i_1 \ldots i_d}(X_{i_1}, \ldots, X_{i_d})$ precisely once. As there are $d(d-1) = (d)_2$ such pairs, we arrive at the result. □

In fact, there are at least two larger families of difference operators which satisfy similar "invariance properties" with respect to the Hoeffding decomposition.

One family of this type can be defined via

$$d_1 := \sum_i D_i, \qquad d_2 := d_1^2 \qquad \text{and more generally} \qquad d_k := d_1^k$$

for any $k \in \{1, 2, \ldots, n\}$. Another one is given by

$$d_k^* := \sum_{i_1 \neq i_2 \neq \ldots \neq i_k} D_{i_1} \ldots D_{i_k}$$

for any $k \in \{1, 2, \ldots, n\}$. It is possible to relate these two families to each other by representing the d_k^* as polynomials in d_1, e. g. we have $d_2^* = d_1^2 - d_1$.

As in the proof of Theorem 3.5, simple combinatorial arguments show that all the d_k and d_k^* operate diagonally on the Hoeffding decomposition. In case of the d_k^*, the eigenvalues of the Hoeffding terms of order up to $k-1$ are 0.

In particular, with Δ as in (4), we see that we have $\Delta = d_2^*$. In other words, Δ is just the second order difference operator from the family of those which annihilate the lower order Hoeffding terms. This corresponds well to our basic concept of second order concentration. Moreover, it parallels the case of the ordinary Laplacian which in particular annihilates linear and affine functions, thus another justification for calling Δ a Laplacian.

An obvious idea is to study concentration of higher order for functions of independent random variables with the help of the operators d_k^*. It seems that this will get more involved than in the second order case, and we intend to study it more in detail in future research.

4 Modified Logarithmic Sobolev Inequalities

A crucial tool in our work is a modified version of the logarithmic Sobolev inequality (LSI) adapted to the difference operators we introduced in the previous section.

For that, we first recall the notion of *entropy.* Let μ be a probability measure on some measurable space $(X, \mathcal{X})$ and $g\colon X \to [0, \infty)$ a measurable function. Then, we define the entropy of g with respect to μ by

$$\mathrm{Ent}(g) := \mathrm{Ent}_\mu(g) := \int g \log g d\mu - \int g d\mu \log \int g d\mu.$$

Here, we set $\mathrm{Ent}(g) := \infty$ if any of the integrals involved does not exist. A common condition for this is whether the integral of $g \log(1+g)$ is finite or not. It is well-known that $\mathrm{Ent}(g) \in [0, \infty]$, which can be shown by applying Jensen's inequality to the function $x \mapsto x \log(x)$. Now we are ready to define

Definition 4.1. *Let μ be a probability measure on some measurable space $(X, \mathcal{X})$, and let Γ be a difference operator on this space satisfying Conditions 3.1. Then, μ satisfies a modified logarithmic Sobolev inequality with constant $\sigma^2 > 0$ with respect to Γ if for any measurable function $f\colon X \to \mathbb{R}$ such that the following integrals are finite we have*

$$\mathrm{Ent}(e^f) \leq \frac{\sigma^2}{2} \int |\Gamma f|^2 e^f d\mu. \tag{7}$$

Here, $|\Gamma f|$ denotes the Euclidean norm of the gradient Γf (which we extend to probability measures on an arbitrary measurable space in this context).

This definition goes back to [B-G1], where it is called LSI_{σ^2}. The term "modified logarithmic Sobolev inequality" is due to Ledoux [L3], Chapter 5.3, where other modifications of logarithmic Sobolev inequalities are discussed as well. The difference between the usual form of the LSI and modified one in (7) is motivated by the fact that the difference operators from Example 3.2 do not satify any sort of chain rule.

The number $\sigma^2 > 0$ is also called *Sobolev constant.* If we occasionally do not consider the Sobolev constant σ^2 itself but its root σ, we will always assume it to be positive as well.

Now we want to relate Definition 4.1 to the gradient operators we introduced in Example 3.2. Unlike in the previous section, using D in this context would cause inconveniences. In this case, only discrete probability measures with a finite number of atoms would have a chance to fulfill a modified LSI of type (7), and the Sobolev constant σ^2 would soon turn pretty bad (depending on the smallest non-zero value among the probabilities of the various atoms). The background is that already in case of two-point measures with atom probabilities p and $1-p$, the constant σ^2 will tend to ∞ if $p \to 0$ or $p \to 1$, and we can "embed" such two-point spaces in any space with a larger amount of atoms or with a continuous distribution, for instance.

This is different in case of ∇. Here, we have the following:

Proposition 4.2. *Let μ be any probability measure on some measurable space $(X, \mathcal{X})$. Then, μ satisfies the modified LSI* (7) *with Sobolev constant $\sigma^2 = 2$ with respect to ∇. Here, ∇ is the gradient operator from Example 3.2.2 (which we extend to probability measures on an arbitrary measurable space in this context).*

Proof. This is due to [B-G2], whose arguments we briefly recall. Noting that we only need (2) in dimension one in the present situation, we apply Jensen's inequality to get

$$\begin{aligned}\mathrm{Ent}_\mu(e^g) \le \mathrm{Cov}_\mu(g, e^g) &= \frac{1}{2}\iint (g(x)-g(y))(e^{g(x)}-e^{g(y)})\mu(dx)\mu(dy)\\ &\le \frac{1}{4}\iint (g(x)-g(y))^2(e^{g(x)}+e^{g(y)})\mu(dx)\mu(dy)\\ &= \int |\nabla g|^2 e^g d\mu.\end{aligned}$$

Here g is any real-valued measurable function on X such that the integrals involved are finite, and the next-to-last step uses the elementary estimate $(a-b)(e^a - e^b) \le \frac{1}{2}(a-b)^2(e^a + e^b)$ for all $a, b \in \mathbb{R}$. However, this means that μ satisfies the modified LSI (7) with Sobolev constant $\sigma^2 = 2$. □

If we especially consider two-point measures, the Sobolev constant can still be improved a little as we see next:

Proposition 4.3. *Let $\mu = p\delta_{+1} + (1-p)\delta_{-1}$ for some $p \in (0,1)$, where δ_x denotes the Dirac measure in $x \in \mathbb{R}$. Then, μ satisfies the modified LSI* (7) *with Sobolev constant $\sigma^2 = 1$ with respect to ∇ as in Example 3.2.2.*

This is again due to [B-G2], and we omit the proof here. It is easy to verify that for instance in case of the symmetric Bernoulli measure (i. e. $p = \frac{1}{2}$ in the situation we described above), this constant is optimal.

We now go on to product spaces. As it is well-known, if we consider n probability spaces each satisfying a logarithmic Sobolev inequality in the usual sense, then the product space will also do so. The same holds for the modified LSI (7). This is a consequence of the following lemma:

Lemma 4.4. *For all $i = 1, \ldots, n$, let $(X_i, \mathcal{X}_i)$ be measurable spaces equipped with probability measures μ_i each satisfying the modified LSI* (7) *with Sobolev constants $\sigma_i^2 > 0$ with respect to ∇ as in Example 3.2.2. Then, the product measure $\mu_1 \otimes \ldots \otimes \mu_n$ on $(X_1 \times \ldots \times X_n, \mathcal{X}_1 \otimes \ldots \otimes \mathcal{X}_n)$ also satisfies the modified LSI* (7) *with Sobolev constant $\sigma^2 = \max_{i=1,\ldots,n} \sigma_i^2$ with respect to ∇.*

Proof. We make use of the fact that in the present situation, the relation

$$\mathrm{Ent}_{\mu_1\otimes\ldots\otimes\mu_n}(g) \leq \sum_{i=1}^{n} \int \mathrm{Ent}_{\mu_i}(g_i)d\mu_1 \otimes \ldots \otimes \mu_n$$

holds. Here, g is an arbitrary non-negative function on $X_1 \times \ldots \times X_n$ and g_i its i-th intersection $g_i(x_i) := g(x_1, \ldots, x_i, \ldots, x_n)$ (with $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n$ fixed) (cf. [L3], Proposition 5.6).

We can assume that $n = 2$ (the rest follows by induction). So let f be a measurable function on $X_1 \times X_2$. Kept in mind that the intersections fulfill $(e^f)_1 = e^{f_1}$ and $(e^f)_2 = e^{f_2}$, respectively, and also that we have $\nabla f_1 = \nabla_1 f$, we obtain by using Fubini that

$$\begin{aligned}
\mathrm{Ent}_{\mu_1\otimes\mu_2}(e^f) &\leq \int \mathrm{Ent}_{\mu_1}(e^{f_1})d\mu_1 \otimes \mu_2 + \int \mathrm{Ent}_{\mu_2}(e^{f_2})d\mu_1 \otimes \mu_2 \\
&\leq \iint \frac{\sigma_1^2}{2}|\nabla f_1|^2 e^{f_1} d\mu_1 d\mu_2 + \iint \frac{\sigma_2^2}{2}|\nabla f_2|^2 e^{f_2} d\mu_2 d\mu_1 \\
&\leq \iint \frac{\max(\sigma_1^2, \sigma_2^2)}{2}(|\nabla_1 f|^2 + |\nabla_2 f|^2)e^f d\mu_1 d\mu_2 \\
&= \frac{\max(\sigma_1^2, \sigma_2^2)}{2} \int |\nabla f|^2 e^f d\mu_1 \otimes \mu_2,
\end{aligned}$$

which completes the proof. □

Therefore, Propositions 4.2 and 4.3 naturally extend to product measures, and we get that any product of such measures satisfies the modified LSI (7) with Sobolev constants 2 and 1, respectively.

In the next section, we will see that for technical reasons we also need modified LSI results for ∇^+. The following lemma enables us to adapt the properties we already proved for ∇. Again, for the sake of generality we extend the definition of ∇^+ from Example 3.2.3 to arbitrary measurable spaces $(X, \mathcal{X})$ and products of them.

Lemma 4.5. *For all $i = 1, \ldots, n$, let $(X_i, \mathcal{X}_i)$ be measurable spaces equipped with probability measures μ_i such that the product measure $\mu_1 \otimes \ldots \otimes \mu_n$ on $(X_1 \times \ldots \times X_n, \mathcal{X}_1 \otimes \ldots \otimes \mathcal{X}_n)$ satisfies the modified LSI* (7) *with Sobolev constant $\sigma^2 > 0$ with respect to ∇ as in Example 3.2.2. Then, $\mu_1 \otimes \ldots \otimes \mu_n$ also satisfies the modified LSI* (7) *with respect to ∇^+ as in Example 3.2.3, and the Sobolev constant can be chosen $2\sigma^2$.*

Proof. We only need to prove this in one dimension. So let g be any real-valued measurable function such that the following integrals are finite. Then, we have

$$\int |\nabla g|^2 e^g d\mu = \frac{1}{2} \iint (g(x) - g(y))^2 e^{g(x)} \mu(dx)\mu(dy)$$

$$
\begin{aligned}
&= \frac{1}{2}\iint \big((g(x)-g(y))_+^2 e^{g(x)} + (g(y)-g(x))_+^2 e^{g(x)}\big)\mu(dx)\mu(dy)\\
&\le \frac{1}{2}\iint \big((g(x)-g(y))_+^2 e^{g(x)} + (g(y)-g(x))_+^2 e^{g(y)}\big)\mu(dx)\mu(dy)\\
&= \frac{1}{2}\Big(2\iint (g(x)-g(y))_+^2 e^{g(x)}\mu(dx)\mu(dy)\Big)\\
&= 2\int |\nabla^+ g|^2 e^g d\mu,
\end{aligned}
$$

where we only used the monotonicity of the exponential function and Fubini's theorem. This completes the proof. □

As a result, we can transport Propositions 4.2 and 4.3 and Lemma 4.4 to the ∇^+ situation. For facilitating references, we explicitly state this in the following proposition:

Proposition 4.6. *For all $i = 1, \ldots, n$, let $(X_i, \mathcal{X}_i)$ be measurable spaces equipped with probability measures μ_i. Then, the product measure $\mu_1 \otimes \ldots \otimes \mu_n$ on $(X_1 \times \ldots \times X_n, \mathcal{X}_1 \otimes \ldots \otimes \mathcal{X}_n)$ satisfies the modified LSI* (7) *with Sobolev constant $\sigma^2 = 4$ with respect to ∇^+ as in Example 3.2.3. If all the X_i are two-point spaces we can take $\sigma^2 = 2$.*

5 Exponential Inequalities

In this section, we derive an inequality for exponential moments similar to [B-C-G], Proposition 2.1. The latter result is as follows: Let (M, d) be a metric space, equipped with some Borel probability measure μ which satisfies a logarithmic Sobolev inequality with constant σ^2. Then, for any locally Lipschitz function f on M such that $\int f d\mu = 0$, $|\nabla|\nabla f|| \le 1$ on the support of μ and $|\nabla f|$ is locally Lipschitz, we have

$$
\int \exp\left(\frac{1}{2\sigma^2} f\right) d\mu \le \exp\left(\frac{1}{2\sigma^2}\int |\nabla f|^2 d\mu\right). \tag{8}
$$

In this context, $|\nabla f|$ and similar expressions always refer to the generalized modulus of the gradient (see [B-C-G]).

This result is based on the iteration of two further inequalities, which are as follows: In the situation sketched above, let $u \colon M \to \mathbb{R}$ is a μ-integrable locally Lipschitz function. Then, we have

$$
\int e^{u - \int u d\mu} d\mu \le \int e^{\sigma^2 |\nabla u|^2} d\mu. \tag{9}
$$

Moreover, if we additionally require $|\nabla u| \le 1$, we have

$$
\int e^{tu^2} d\mu \le \exp\left(\frac{t}{1 - 2\sigma^2 t}\int u^2 d\mu\right) \tag{10}
$$

for any $0 \leq t < \frac{1}{2\sigma^2}$. Inequality (8) then follows by first applying (9) to u and then (10) with u replaced by $|\nabla u|$.

In the sequel, we derive analogous results for functions of independent random variables. Note that anything we do in this section holds for an arbitrary difference operator satisfying Conditions 3.1. Due to the results from the previous section we will always take either ∇ or ∇^+ from Example 3.2, however.

So now consider any probability measure on some measurable space $(X, \mathcal{X})$ which satisfies the modified LSI (7) with Sobolev constant $\sigma^2 > 0$ with respect to the gradient operator ∇. In Bobkov and Götze [B-G1], it was proved that for all bounded measurable functions $f\colon X \to \mathbb{R}$ such that $\int f d\mu = 0$, we have

$$\int e^f d\mu \leq \int e^{\sigma^2 |\nabla f|^2} d\mu. \tag{11}$$

This is an exact analogue of (9). Its proof is similar to the proof of inequality (14) which will be sketched in the proof of Lemma 5.1.

Note that we did not make a difference between the one-dimensional case and the multi-dimensional one (so, product or non-product spaces) above as in fact, it does not matter in which situation we are. The same holds for the rest of this section, and in most cases we will not explicitly stress this in the sequel.

The adaption of (10) requires more effort. The main problem is that as we already noted, all types of gradient operators we introduced in Example 3.2 do not satisfy any sort of "chain rule".

To illustrate the situation, let for a moment ∇ denote the usual gradient (as opposed to the difference operator from Example 3.2.2) and $|\nabla f|$ its Euclidean norm. Then, if we assume that $|\nabla f| \leq 1$, we immediately get

$$|\nabla f^2| = 2|f||\nabla f| \leq 2|f|$$

by chain rule. However, if we now take ∇ as in (2) instead, such an inequality cannot be true anymore. To see this, consider the space $\{\pm 1\}^2$ equipped with the product of two symmetric Bernoulli distributions. Remembering Remark 3.3.1, the analogue of the above inequality would then be

$$|\nabla f(x)^2|^2 = \frac{1}{4}((f(x)^2 - f(\sigma_1 x)^2)^2 + (f(x)^2 - f(\sigma_2 x)^2)^2) \leq \alpha |f(x)|^2$$

for some $\alpha > 0$. However, it is immediately clear that we can easily construct examples in which this inequality is obviously wrong (by considering the zeros of f). The same holds if we choose D instead of ∇ again due to Remark 3.3.1.

There are several possibilities of addressing this problem. One solution is to introduce an additional summand as a sort of "compensation". We would then consider functions satisfying an inequality of the form $|\nabla f^2|^2 \leq af^2 + b$ on the support of μ for some constants $a, b \geq 0$. Note that in case of the usual gradient and the situation described above, this holds with $a = 4$ and $b = 0$. Indeed, the following arguments remain valid assuming this inequality.

However, it turns out that if we use the gradient ∇^+ from Example 3.2.3 we can proceed in a more elegant way and circumvent most of the additional work. This is due to the fact that unlike in case of ∇, the zeros of f do not play such a significant role in case of ∇^+ anymore because of taking the positive part in the definition of (3).

The principal argument is as follows: let $f\colon X \to \mathbb{R}$ be any measurable function on some probability space $(X, \mathcal{X}, \mu)$. Then, for any $x, y \in X$ we get

$$\begin{aligned}(f(x)^2 - f(y)^2)_+^2 &= (|f(x)| + |f(y)|)^2(|f(x)| - |f(y)|)_+^2 \\ &\leq 4|f(x)|^2(|f(x)| - |f(y)|)_+^2\end{aligned}$$

as we have $|f(x)| \geq |f(y)|$ for any values x and y such that the two sides of this inequality do not vanish. Taking integrals and roots, we thus get that for any function $f\colon X \to \mathbb{R}$ in $L^2(\mu)$ such that $|\nabla^+|f|| \leq 1$ on the support of μ, we have

$$|\nabla^+ f^2| \leq 2|f|. \tag{12}$$

Note that this also holds for product measures (i. e. the multivariate case), where deriving it works by applying everything we just did componentwise.

However, now we have arrived at the same basic inequality as in case of the usual gradient on $(\mathbb{R}^n, \mathbb{B}^n)$, or at a more general level the deviation case in [B-G1]. We can therefore proceed the same way as in these cases, and using an argument which is basically due to S. Aida, T. Masuda and I. Shikegawa [A-M-S] we get a full analogue of (10) which is as follows:

Lemma 5.1. *Let μ be a probability measure on some measurable space $(X, \mathcal{X})$ which satifies the modified LSI* (7) *with Sobolev constant $\tilde{\sigma}^2 > 0$ with respect to the gradient operator ∇^+ from Example 3.2.3. Moreover, let $f\colon X \to \mathbb{R}$ be a bounded measurable function such that $|\nabla|f|| \leq 1$ on the support of μ. Then we have*

$$\int e^{tf^2} d\mu \leq \exp\left(\frac{t}{1 - 2\tilde{\sigma}^2 t} \int f^2 d\mu\right) \tag{13}$$

for all $t \in [0, \frac{1}{2\tilde{\sigma}^2})$.

At first sight, the fact that we have to use the absolute value of f in the condition $|\nabla|f|| \leq 1$ might not seem optimal. However, we will finally apply this inequality to non-negative functions only anyway.

Proof. First, note that by the very definitions of ∇ and ∇^+ as given in Example 3.2, we have $0 \leq \nabla_i^+ f(x) \leq \nabla_i f(x)$ for any function $f \in L^2(\mu)$, all $x \in \mathbb{R}^n$ and all $i = 1, \ldots, n$. In particular, it follows that $|\nabla^+ f(x)| \leq |\nabla f(x)|$. Therefore, the condition $|\nabla|f|| \leq 1$ implies $|\nabla^+|f|| \leq 1$, which is the condition we actually make use of in the sequel.

From now on, we just adapt the arguments from [B-G1], p. 6 f. Their starting point is the inequality

$$\int e^f d\mu \leq \left(\int e^{\lambda f + (1-\lambda)\tilde{\sigma}^2 |\nabla^+ f|^2/2} d\mu \right)^{1/\lambda} \tag{14}$$

for all bounded measurable $f\colon X \to \mathbb{R}$ and all $\lambda \in (0,1]$. Here we have already plugged in ∇^+ as our choice of the difference operator.

To deduce (14), we use that

$$\mathrm{Ent}(g) = \sup \left\{ \int gh\, d\mu\colon h\colon X \to \mathbb{R} \text{ measurable s. th. } \int e^h d\mu \leq 1 \right\}.$$

If we set $g := e^f$ und $h := \lambda f + (1-\lambda)\tilde{\sigma}^2 |\nabla^+ f|^2/2 - \beta$ with $\beta = \log \int e^{\lambda f + (1-\lambda)\tilde{\sigma}^2 |\nabla^+ f|^2/2} d\mu$ in this context, we have $\int e^h d\mu = 1$ and thus

$$\int (\lambda f + (1-\lambda)\tilde{\sigma}^2 |\nabla^+ f|^2/2 - \beta) e^f d\mu \leq \mathrm{Ent}(e^f).$$

Since f satisfies the modified LSI (7) with constant $\tilde{\sigma}^2$, it follows that

$$\lambda \int f e^f d\mu + (1-\lambda)\mathrm{Ent}(e^f) - \beta \int e^f d\mu \leq \mathrm{Ent}(e^f).$$

This is equivalent with

$$\lambda \int e^f d\mu \log \int e^f d\mu - \beta \int e^f d\mu \leq 0,$$

from which we directly get (14).

We now apply (14) to the function $sf^2/(2\tilde{\sigma}^2)$ with $0 < s < 1$ and $\lambda = (p-s)/(1-s)$ for any $p \in (s,1]$. Together with (12), this gives us

$$\int e^{sf^2/(2\tilde{\sigma}^2)} d\mu \leq \left(\int \exp\left(\frac{psf^2}{2\tilde{\sigma}^2} \right) d\mu \right)^{(1-s)/(p-s)}.$$

For $p = 1$ both sides are equal, and as for $p < 1$ the upper inequality holds, we get that the logarithm of the left hand side (considered as a function of p) must increase more rapidly at $p = 1$ than that of the right hand side. We thus consider the derivatives of the logarithms of both sides at $p = 1$ and arrive at the inequality

$$0 \geq \frac{1}{1-s} \left[(1-s) \int \frac{sf^2}{2\tilde{\sigma}^2} e^{sf^2/(2\tilde{\sigma}^2)} d\mu - \int e^{sf^2/(2\tilde{\sigma}^2)} d\mu \log \int e^{sf^2/(2\tilde{\sigma}^2)} d\mu \right].$$

Now we set

$$u(s) := \int e^{sf^2/(2\tilde{\sigma}^2)} d\mu,$$

$s \in (0,1]$. Then we get

$$0 \geq \frac{1}{1-s}\left[s(1-s)u'(s) - u(s)\log u(s)\right],$$

or equivalently

$$0 \geq \frac{1-s}{s}\frac{u'(s)}{u(s)} - \frac{1}{s^2}\log u(s).$$

Hence, the function

$$v(s) := \exp\left(\frac{1-s}{s}\log u(s)\right)$$

is non-increasing in s, and therefore we have $v(s) \leq \lim_{s\downarrow 0} v(s) =: v(0^+)$ for all $s \in (0,1]$.

Note that

$$\begin{aligned} v(0^+) &= \lim_{s\downarrow 0}\left(u(s)^{(1-s)/s}\right) = \lim_{s\downarrow 0}\left(\int e^{sf^2/(2\tilde{\sigma}^2)}d\mu\right)^{(1-s)/s} \\ &= \exp\left(\frac{1}{2\tilde{\sigma}^2}\int f^2 d\mu\right). \end{aligned}$$

Thus, we have

$$\exp\left(\frac{1-s}{s}\log u(s)\right) \leq \exp\left(\frac{1}{2\tilde{\sigma}^2}\int f^2 d\mu\right)$$

for all $s \in (0,1]$, or equivalently

$$\int e^{sf^2/(2\tilde{\sigma}^2)}d\mu \leq \exp\left(\frac{1}{2\tilde{\sigma}^2}\frac{s}{1-s}\int f^2 d\mu\right).$$

Setting $t = s/(2\tilde{\sigma}^2)$ completes the proof. □

Combining inequalities (11) and (13), we now get the following analogue of Proposition 2.1 from [B-C-G]:

Proposition 5.2. *Let μ be a probability measure on some measurable space $(X, \mathcal{X})$ which satisfies the modified LSI (7) with Sobolev constant $\sigma^2 > 0$ with respect to the gradient operator ∇ and which moreover satisfies the modified LSI (7) with Sobolev constant $\tilde{\sigma}^2$ with respect to the gradient operator ∇^+. Furthermore, let $f\colon X \to \mathbb{R}$ be a bounded measurable function such that $\int f d\mu = 0$ and $|\nabla|\nabla f|| \leq 1$ on the support of μ. Then, we have*

$$\int \exp\left(\frac{1}{2\sigma\tilde{\sigma}}f\right)d\mu \leq \exp\left(\frac{1}{2\tilde{\sigma}^2}\int |\nabla f|^2 d\mu\right).$$

It might seem cumbersome to require μ to satisfy even two modified LSIs with respect to two different gradients. However, due to our results from the previous section these conditions are merely of formal nature.

Proof. First, applying (11) to λf leads to

$$\int e^{\lambda f} d\mu \leq \int e^{\lambda^2 \sigma^2 |\nabla f|^2} d\mu.$$

Moreover, (13) with $t = \lambda^2 \sigma^2$ for any $\lambda \in [0, \frac{1}{\sqrt{2}\sigma\tilde{\sigma}})$ and with f replaced by $|\nabla f|$ gives us

$$\int e^{\lambda^2 \sigma^2 |\nabla f|^2} d\mu \leq \exp\left(\frac{\lambda^2 \sigma^2}{1 - 2\sigma^2 \tilde{\sigma}^2 \lambda^2} \int |\nabla f|^2 d\mu\right).$$

Combining these two inequalities then yields

$$\int e^{\lambda f} d\mu \leq \exp\left(\frac{\lambda^2 \sigma^2}{1 - 2\sigma^2 \tilde{\sigma}^2 \lambda^2} \int |\nabla f|^2 d\mu\right).$$

Setting $\lambda = \frac{1}{2\sigma\tilde{\sigma}}$ completes the proof. □

6 Relating First Order and Second Order Difference Operators

In Proposition 5.2, we estimated the exponential moments of f involving a first order difference operator. Now we go on to second order difference operators. Here, it is convenient to work with the operator D as defined in (1) and to make use of the Laplacian as introduced in (4), including properties like Theorem 3.5.

As in [B-C-G], we will consider second order differences in form of a suitably defined "Hessian". Remember we have $D_{ii} f = D_i f$ as discussed in the context of (4). Therefore, taking the $n \times n$-matrix with entries $(f''(x))_{ij} = D_{ij} f(x)$ would lead to first order differences on the diagonal. For this reason, we set

$$f^{\hat{\prime\prime}}(x)_{ij} := \begin{cases} D_{ij} f(x), & i \neq j, \\ 0, & i = j, \end{cases} \tag{15}$$

instead. In other words, we take the Hessian of f with respect to the difference operator D but remove its diagonal. Note that this definition also corresponds well to the Laplacian (4).

Our next aim is to derive an inequality of the form

$$\gamma \mathbb{E} |\nabla f|^2 \leq \mathbb{E} \|f^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2$$

for some constant $\gamma > 0$. Here, as before, $\mathbb{E}$ denotes taking the expectation with respect to the product measure $\mu = \otimes_{i=1}^n \mu_i$. One of our main tools is the following lemma:

Lemma 6.1. *Let $\mu_1, \ldots, \mu_n$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and denote by $\mu = \otimes_{i=1}^n \mu_i$ their product measure. Consider $f, g \in L^2(\mu)$. Then, we have:*

1. *$\mathbb{E}(D_i f)g = \mathbb{E}f(D_i g) = \mathbb{E}(D_i f)(D_i g)$, where D_i is the difference operator from Example 3.2.1.*
2. *$\mathbb{E}(Df)g = \mathbb{E}f(Dg)$, where D is the gradient operator from Example 3.2.1 and the integral has to be understood componentwise.*
3. *$\mathbb{E}(\Delta f)g = \mathbb{E}f(\Delta g) = \sum_{i \neq j} \mathbb{E}(D_{ij} f)(D_{ij} g)$, where Δ is the Laplacian as in (4).*

Hence, the difference operators D_i, the gradient operator D and the Laplacian Δ are in some sense selfadjoint operators on $L^2(\mu)$.

Proof. Part 1 follows from the fact that by Fubini's theorem, we have

$$\mathbb{E}g\mathbb{E}_i f = \mathbb{E}\mathbb{E}_i f \mathbb{E}_i g = \mathbb{E}f\mathbb{E}_i g.$$

Here, $\mathbb{E}_i$ once again denotes taking expected values in the i-th component. Having established Part 1, Parts 2 and 3 are immediate. Note that in Part 3 we use that we always have $D_{ij} f = D_{ji} f$ for any i, j by (1) and Fubini's theorem. □

Using this result, we can prove an inequality of the form in question:

Proposition 6.2. *Let $\mu_1, \ldots, \mu_n$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and denote by $\mu = \otimes_{i=1}^n \mu_i$ their product measure. Let $f \in L^2(\mu)$ be a function such that its Hoeffding decomposition with respect to μ is given by*

$$f = \sum_{k=d}^{n} f_k$$

for some $d \geq 2$. Then we have

$$\int |\nabla f|^2 d\mu \leq \frac{1}{d-1} \int \|f^{\hat{''}}\|_{\mathrm{HS}}^2 d\mu.$$

Equality holds if we have $f = f_d$, i. e. the Hoeffding decomposition of f consists of a single term only. Here, $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert Schmidt norm of a matrix.

Proof. First, let $f = f_d$. Then, we have due to Lemma 6.1.3 (using the notation introduced there)

$$\mathbb{E}\|f^{\hat{''}}\|_{\mathrm{HS}}^2 = \sum_{i \neq j} \mathbb{E}(D_{ij} f)(D_{ij} f) = \mathbb{E}f\Delta f.$$

Moreover, Theorem 3.5 yields $\Delta f = (d)_2 f$ (since $f = f_d$). Consequently, we have

$$\mathbb{E}\|f^{\hat{''}}\|_{\mathrm{HS}}^2 = (d)_2 \mathbb{E}f^2. \qquad (*)$$

On the other hand, we clearly have $|\nabla f|^2 = (\nabla_1 f)^2 + \ldots + (\nabla_n f)^2$. Moreover, if $X_1, \ldots, X_n$ is a set of independent random variables with distributions μ_i,

$i = 1, \dots, n$, we have $f_d(X_1, \dots, X_n) = \sum_{i_1 < \dots < i_d} h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})$, where the summands on the right hand side are pairwise orthogonal in L^2. Here we used the notation from the proof of Theorem 3.5.

Now let $\bar{X}_1, \dots, \bar{X}_n$ be a set of independent copies of the random variables $X_1, \dots, X_n$ (cf. Remark 3.3.3). We then extend the family of the $h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})$ by considering the same functions but we replace precisely one entry X_{i_j} by the corresponding independent copy $\bar{X}_{i_j}$ (in other words, and once again using the notation from Remark 3.3.3, we take the $T_{i_j} h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})$). Doing so, we still have a (larger) family of pairwise orthogonal functions in L^2

$$\bigcup_{i_1 < \dots < i_d} \{h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})\} \cup \{T_{i_j} h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d}), j = 1, \dots, d\},$$

now integrating with respect to the X_i and the $\bar{X}_i$, however.

We easily see that still using Remark 3.3.3, we get (similarly to the deduction of (5))

$$\begin{aligned}(\nabla_i f_d(X_1, \dots, X_n))^2 &= \frac{1}{2} \bar{\mathbb{E}}_i (f_d - T_i f_d)^2 \\ =& \frac{1}{2} \bar{\mathbb{E}}_i \Big(\sum_{\substack{i_1 < \dots < i_d \\ i \in \{i_1, \dots, i_d\}}} (h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d}) - T_i h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})) \Big)^2 .\end{aligned}$$

Together with the orthogonality relations as described above, we get (once again applying Fubini's theorem)

$$\begin{aligned}&\mathbb{E}(\nabla_i f_d(X_1, \dots, X_n))^2 \\ &= \sum_{\substack{i_1 < \dots < i_d \\ i \in \{i_1, \dots, i_d\}}} \frac{1}{2} \Big(\mathbb{E}\bar{\mathbb{E}}_i (h^2_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d}) + T_i h^2_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})) \Big) \\ &= \sum_{\substack{i_1 < \dots < i_d \\ i \in \{i_1, \dots, i_d\}}} \mathbb{E} h^2_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d}).\end{aligned}$$

As in the proof of Theorem 3.5, it therefore remains to check how often each term $\mathbb{E} h^2_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})$ appears in $\mathbb{E}|\nabla f|^2 = \sum_i (\nabla_i f_d)^2$. However, it is clear that each $i \in \{i_1, \dots, i_d\}$ replicates the summand $\mathbb{E} h_{i_1 \dots i_d}(X_{i_1}, \dots, X_{i_d})$ exactly once. Consequently, it follows that

$$\mathbb{E}|\nabla f|^2 = d \mathbb{E} f^2. \qquad (**)$$

Comparing $(*)$ and $(**)$ completes the proof in case of $f = f_d$.

In the general case we make use of the orthogonality of the Hoeffding decomposition and get

$$\mathbb{E}|\nabla f|^2 = \sum_{k=d}^{n} \frac{1}{k-1} \mathbb{E} \| f_k^{\hat{\prime\prime}} \|^2_{\mathrm{HS}} \leq \frac{1}{d-1} \mathbb{E} \| f^{\hat{\prime\prime}} \|^2_{\mathrm{HS}}.$$

Here we used that the integrals of the Hilbert Schmidt norms of the second order differences of the Hoeffding terms sum up for reasons of orthogonality. This finally completes the proof. □

Note that in Proposition 5.2, the condition $|\nabla|\nabla f|| \leq 1$ also involves second order difference operators. It would be desirable to replace it by a simpler condition, for instance involving a Hessian again. However, we have not found a satisfactory solution for this problem.

We are now ready to prove Theorems 1.1 and 1.2:

Proof of Theorem 1.1 and Theorem 1.2. First, combine Proposition 5.2 with $(X, \mathcal{X}) = (\mathbb{R}^n, \mathbb{B}^n)$ and Proposition 6.2. This leads us to

$$\int \exp\left(\frac{1}{2\sigma\tilde{\sigma}} f\right) d\mu \leq \exp\left(\frac{1}{2\tilde{\sigma}^2} \frac{1}{d-1} \int \|f^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mu\right) \tag{16}$$

if μ satisfies the modified LSI (7) with constant $\sigma^2 > 0$ with respect to ∇ and furthermore with constant $\tilde{\sigma}^2 > 0$ with respect to ∇^+.

Now, from (16) we get

$$\begin{aligned}\int \exp\left(\frac{1}{2\sigma\tilde{\sigma}} |f|\right) d\mu &\leq \int \left(\exp\left(\frac{1}{2\sigma\tilde{\sigma}} f\right) + \exp\left(\frac{1}{2\sigma\tilde{\sigma}} (-f)\right)\right) d\mu \\ &\leq 2\exp\left(\frac{1}{2\tilde{\sigma}^2} \frac{1}{d-1} \int \|f^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mu\right). \end{aligned} \tag{17}$$

Thus, by applying Hölder's inequality we obtain

$$\int \exp\left(\frac{1}{2\sigma\tilde{\sigma}\kappa} |f|\right) d\mu \leq \left(2\exp\left(\frac{1}{2\tilde{\sigma}^2} \frac{1}{d-1} \int \|f^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mu\right)\right)^{1/\kappa}$$

for all $\kappa \geq 1$. Using $\int \|f^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mu \leq b^2$, we see that this is ≤ 2 if

$$\kappa \geq \left(\log 2 + \frac{1}{2\tilde{\sigma}^2} b^2/(d-1)\right) / \log 2, \qquad (*)$$

or equivalently

$$\frac{1}{2\sigma\tilde{\sigma}\kappa} \leq \frac{\log 2}{2\sigma\tilde{\sigma}\log 2 + \frac{\sigma}{\tilde{\sigma}} b^2/(d-1)}.$$

Note that from $(*)$ we immediately get that any such κ will be ≥ 1 as required.

From Proposition 4.2, Proposition 4.3, Lemma 4.4 and Proposition 4.6 we know that μ indeed satisfies modified LSIs of type (7) both with respect to ∇ and with respect to ∇^+, and that we can set $\sigma^2 = 2$ and $\tilde{\sigma}^2 = 4$ or, in the Bernoulli case, $\sigma^2 = 1$ and $\tilde{\sigma}^2 = 2$. We thus get

$$\int \exp\left(\frac{\log 2}{\sqrt{32}\log 2 + \frac{1}{\sqrt{2}} b^2/(d-1)} |f|\right) d\mu \leq 2$$

if $\sigma^2 = 2$ and $\tilde{\sigma}^2 = 4$ and

$$\int \exp\left(\frac{\log 2}{\sqrt{8}\log 2 + \frac{1}{\sqrt{2}}b^2/(d-1)}|f|\right) d\mu \leq 2$$

if $\sigma^2 = 1$ and $\tilde{\sigma}^2 = 2$. Noting that

$$\frac{\log 2}{\sqrt{32}\log 2 + \frac{1}{\sqrt{2}}x} \geq \frac{1}{6+2x} \qquad \text{and} \qquad \frac{\log 2}{\sqrt{8}\log 2 + \frac{1}{\sqrt{2}}x} \geq \frac{1}{3+2x}$$

for all $x \geq 0$ completes the proof. □

7 Allowing First Order Hoeffding Terms

In this section, we give a version of Theorem 1.1 which also allows functions with non-vanishing first order Hoeffding terms. This is an analogue of Theorem 1.3 in Bobkov, Chistyakov and Götze [B-C-G] and will be done by setting up conditions which guarantee that the limit behavior as $n \to \infty$ does not change. However, establishing this needs some extra work.

First we introduce some notation. Let $X_1, \ldots, X_n$ be independent random variables and $f\colon \mathbb{R}^n \to \mathbb{R}$ a function such that $f(X_1, \ldots, X_n)$ has Hoeffding decomposition $f = \sum_{k=0}^n f_k$. Then, we denote by

$$Rf := f - f_0 - f_1 = \sum_{k=2}^n f_k$$

the projection of f onto the space of the statistics $f(X_1, \ldots, X_n)$ whose Hoeffding terms of orders 0 and 1 vanish. Due to (6) it is clear that we have

$$D_{ij}Rf = D_{ij}f \tag{18}$$

for all $i \neq j$.

We always consider statistics with expected value 0 (and therefore $f_0 = 0$). Moreover, as in Section 3, we denote the first order Hoeffding term by

$$f_1(X_1, \ldots, X_n) = \sum_{i=1}^n h_i(X_i).$$

If we want to obtain a result similar to Theorem 1.1, for instance

$$\int e^{c|f|} d\mu \leq \int e^{c(|f_1|+|Rf|)} d\mu \leq 2$$

for some constant $c > 0$, we have set up conditions ensuring $f_1 = \mathcal{O}(1)$. The following theorem presents two ways of doing so:

Theorem 7.1. *Let $\mu_1, \ldots, \mu_n$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and denote by $\mu = \otimes_{i=1}^n \mu_i$ their product measure. Moreover, let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a bounded measurable function such that its Hoeffding decomposition with respect to μ is given by*

$$f = f_1 + \sum_{k=d}^{n} f_k = f_1 + Rf$$

for some $d \geq 2$. (In particular, we have $\mathbb{E}f = 0$.) Denote by D and ∇ the difference operators as introduced in Example 3.2. Suppose that the condition

$$|\nabla|\nabla Rf|| \leq 1$$

is satisfied on the support of μ and that we have

$$\int \|f^{\hat{''}}\|_{\mathrm{HS}}^2 d\mu \leq b^2$$

for some $b \geq 0$. Here, $\|f^{\hat{''}}\|_{\mathrm{HS}}$ is the "dediagonalized" Hessian of f with respect to D, and $\|f^{\hat{''}}\|_{\mathrm{HS}}$ denotes its Hilbert Schmidt norm. Furthermore, assume that one of the conditions

(i) $|\nabla f_1|^2 \leq \gamma^2$ *on the support of μ for some $\gamma \geq 0$ or*

(ii) $|\nabla|\nabla f_1|| \leq 1$ *on the support of μ and $\int |\nabla f_1|^2 d\mu \leq \alpha^2$ for some $\alpha^2 \geq 0$*

is satisfied. Then, there we have

$$\int \exp\left(\frac{1}{12 + 4b^2/(d-1) + 7\gamma}|f|\right) d\mu \leq 2$$

in case of condition (i) and

$$\int \exp\left(\frac{1}{4(3 + b^2/(d-1) + \alpha^2)}|f|\right) d\mu \leq 2$$

in case of condition (ii).

Proof. The basic argument is as follows: if we have two functions φ_1 and φ_2 on $\mathbb{R}^n$ both satisfying

$$\int e^{c_i|\varphi_i|} d\mu \leq 2 \qquad (*)$$

for some constants $c_i > 0$, $i = 1, 2$, then we have

$$\begin{aligned}\int e^{\min(c_1,c_2)|\varphi_1+\varphi_2|/2} d\mu &\leq \int e^{c_1|\varphi_1|/2} e^{c_2|\varphi_2|/2} d\mu \\ &\leq \left(\int e^{c_1|\varphi_1|} d\mu\right)^{1/2} \left(\int e^{c_2|\varphi_2|} d\mu\right)^{1/2} \leq 2\end{aligned}$$

due to the Cauchy Schwarz inequality. In our situation, we set $\varphi_1 = f_1$ and $\varphi_2 = Rf$. Hence, we only have to check $(*)$.

In case of Rf, this clear because of Theorem 1.1. In case of f_1 together with condition (ii), we apply Proposition 5.2 and then proceed as in the proof of Theorem 1.1. Using the notation from $(*)$, this leads to

$$c_1 = \frac{1}{6+2\alpha^2} \qquad \text{and} \qquad c_2 = \frac{1}{6+2b^2/(d-1)},$$

so that we can estimate $\min(c_1, c_2)/2$ as stated in the theorem.

It therefore remains to check $(*)$ in the case of f_1 together with condition (i). Here, inequality (11) yields

$$\int e^{\lambda f_1} d\mu \leq \int e^{\sigma^2\lambda^2|\nabla f_1|^2} d\mu \leq e^{\sigma^2\lambda^2\gamma^2}$$

for any $\lambda > 0$, thus

$$\int e^{\lambda|f_1|} d\mu \leq 2e^{\sigma^2\lambda^2\gamma^2}.$$

As in the proof of Theorem 1.1 it follows that

$$\int e^{\lambda|f_1|/\kappa} d\mu \leq \left(2e^{\sigma^2\lambda^2\gamma^2}\right)^{1/\kappa}.$$

for all $\kappa \geq 1$. This is ≤ 2 if

$$\frac{\lambda}{\kappa} \leq \frac{\lambda \log 2}{\log 2 + \lambda^2\sigma^2\gamma^2}.$$

In particular, we see that any such κ must be ≥ 1 indeed (as long as we require $\kappa > 0$). The expression on the right hand side attains a maximum at $\lambda = (\log 2)^{1/2}/(\sigma\gamma)$ whose value is exactly $(\log 2)^{1/2}/(2\sigma\gamma)$. Plugging in $\sigma^2 = 2$ we thus get

$$\frac{1}{2}c_1 = \frac{(\log 2)^{1/2}}{4\sqrt{2}\gamma} \geq \frac{1}{7\gamma}.$$

With c_2 as in the first part of the proof we arrive at the bound given in the theorem. □

Similarly to Theorem 1.2, it is possible to improve the constants c_1 and c_2 if all the underlying measures are Bernoulli distributions. As the result would be pretty lengthy we will skip the details. The same holds for results similar to Remark 1.3.

Unfortunately, the results of Theorem 7.1 are not too satisfactory. This is because of several aspects:

1) In the condition $|\nabla|\nabla Rf|| \leq 1$, we must use Rf instead of f.

2) Condition (i) exclusively makes use of first order differences and requires them to be pointwise bounded, which does not fit well into the context of this work.

3) In case of condition (ii) we take the expected value of $|\nabla f_1|^2$ instead, however this is at the cost of getting the additional condition $|\nabla|\nabla f_1|| \leq 1$.

These problems do not occur in [B-C-G] in case of the unit sphere. This is because of the following reasons: comparing our results to [B-C-G], the analogue of the first order Hoeffding term f_1 is the "linear part" of a function f in case of the unit sphere. Of course, the linear part has constant derivatives. By contrast, if we apply any of the difference operators from Example 3.2 to f_1, the result will not be constant in general. This complicates controlling the additional terms. It would be desirable to simplify the conditions given in Theorem 7.1 for instance by combining some of them, but it is not clear how this could be achieved.

Finally, we observe that using Example 3.2.2 and Remark 3.3.3, we get

$$\begin{aligned}|\nabla f_1(X_1,\ldots,X_n)|^2 &= \sum_{i=1}^n \frac{1}{2}\bar{\mathbb{E}}_i(h_i(X_i) - h_i(\bar{X}_i))^2 \\ &= \frac{1}{2}\sum_{i=1}^n (h_i(X_i)^2 + \bar{\mathbb{E}}_i h_i(\bar{X}_i)^2)\end{aligned}$$

due to the properties of the Hoeffding decomposition according to Theorem 3.4, and as a consequence

$$\mathbb{E}|\nabla f_1(X_1,\ldots,X_n)|^2 = \sum_{i=1}^n \mathbb{E}h_i(X_i)^2.$$

See also the proof of Theorem 6.2 for similar calculations.

This allows us to replace the condition $|\nabla f_1| = \mathcal{O}(1)$ in (i) and (ii) from Theorem 7.1 by a condition of the form $\max_i |h_i| = \mathcal{O}(1/\sqrt{n})$. This can be especially useful in case of symmetric Bernoulli distributions, as here we have $h_i(X_i) = r_i X_i$ for some $r_i \in \mathbb{R}$ and we therefore only have to bound the numbers r_i.

8 Functions in Independent Bernoulli Variables

To start with the applications of Theorem 1.1, we consider functions of n independent symmetric Bernoulli variables each taking values in $\{\pm 1\}$. It is well-known that a function f of this type can be represented as

$$f(X_1,\ldots,X_n) = \alpha_0 + \sum_{i=1}^n \alpha_i X_i + \sum_{i<j} \alpha_{ij} X_i X_j + \ldots, \tag{19}$$

where the coefficients α_I (with a suitable multi-index I) are real numbers and the sum goes up to order n. More precisely, we have

$$\alpha_{i_1 \ldots i_d} = \mathbb{E}f(X_1,\ldots,X_n)X_{i_1}\cdots X_{i_d}$$

for any $i_1 < \ldots < i_d$, $d = 0, 1, \ldots, n$.

This representation is called the *Fourier-Walsh expansion* of the function f, and the expression on the right-hand side of (19) is also known as a *Rademacher chaos*. It is immediately clear that (19) is at the same time the Hoeffding decomposition of f. Consequently, we see that for applying our results like Theorem 1.2 we need to require (19) to start with the second order terms.

To begin, we will therefore discuss functions of the form

$$f\colon \mathbb{R}^n \to \mathbb{R}; \qquad f(x_1, \ldots, x_n) := \sum_{i<j} \alpha_{ij} x_i x_j \qquad (*)$$

with the α_{ij} being real numbers and the probability measures μ_i all given by the symmetric Bernoulli distribution on $\{\pm 1\}$. Hence, the product measure which we will again denote μ is just the uniform distribution on $\{\pm 1\}^n$.

We then get

$$\int f d\mu = 0 \qquad \text{and} \qquad \int f^2 d\mu = \sum_{i<j} \alpha_{ij}^2, \tag{20}$$

as we readily check. From the proof of Proposition 6.2, we therefore obtain that due to $f = f_2$ (using the notation from Theorem 3.4), we have

$$\int \|f^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mu = (2)_2 \int f^2 d\mu = 2 \sum_{i<j} \alpha_{ij}^2. \tag{21}$$

So what remains to be checked is the condition $|\nabla|\nabla f|| \leq 1$ for all $x \in \{\pm 1\}^n$. To simplify notation, we introduce the convention that $\sum^{(j)}$ means summing over all indexes but j. Similarly, $\sum^{(j,k)}$ denotes summing over all indexes but j and k. Furthermore, in the sequel we will assume that $\alpha_{ij} = \alpha_{ji}$ for all $i > j$ (note that a priori we only allowed indexes $i < j$).

With the help of Remark 3.3.1 (in particular, remember the operator σ_k), we can rewrite

$$|\nabla|\nabla f(x)|| = \frac{1}{2}\Big(\sum_{k=1}^n (|\nabla f(x)| - |\nabla f(\sigma_k x)|)^2\Big)^{1/2}. \tag{22}$$

We next estimate the difference of the two norms in (22).

For this, once again due to Remark 3.3.1, we observe that for a function f as in $(*)$, we have

$$(\nabla_i f(x))^2 = (D_i f(x))^2 = \big(\sum_{j=1}^n{}^{(i)} \alpha_{ij} x_i x_j\big)^2 = \big(\sum_{j=1}^n{}^{(i)} \alpha_{ij} x_j\big)^2$$

(using (5) in the second step as we have $f = f_2$). If $i \neq k$, this becomes

$$\big(\sum_{j=1}^n{}^{(i,k)} \alpha_{ij} x_j + \alpha_{ik} x_k\big)^2.$$

Likewise, assuming $i \neq k$ we have

$$(\nabla_i f(\sigma_k x))^2 = \big(\sum_{j=1}^{n}{}^{(i,k)}\alpha_{ij}x_j - \alpha_{ik}x_k\big)^2.$$

If $i = k$, we have $(\nabla_k f(x))^2 = (\nabla_k f(\sigma_k x))^2$.

This means that if we set v to be the vector with entries $\sum_{j=1}^{n}{}^{(i,k)}\alpha_{ij}x_j$ for all $i \neq k$ and entry $\sum_{j=1}^{n}{}^{(k)}\alpha_{kj}x_j$ in the k-th component and moreover w to be the vector with entries $\alpha_{ik}x_k$ for all $i \neq k$ and k-th component $w_k = 0$, we have

$$\begin{aligned}||\nabla f(x)| - |\nabla f(\sigma_k(x))|| &= ||v + w| - |v - w|| \\ &\leq 2|w| = 2\big(\sum_{i=1}^{n}{}^{(k)}\alpha_{ik}^2\big)^{1/2}.\end{aligned}$$

Going back to (22) and plugging in leads to

$$|\nabla|\nabla f(x)|| \leq (\sum_{i\neq k}\alpha_{ik}^2)^{1/2} = \sqrt{2}(\sum_{i<j}\alpha_{ij}^2)^{1/2}. \tag{23}$$

The term on the right-hand side equals $\|\hat{f''}(x)\|_{\mathrm{HS}}$, which is the same as $\mathbb{E}\|\hat{f''}(x)\|_{\mathrm{HS}}$ in this case.

We now apply Remark 1.3. For this, set $A := (\sum_{i<j}\alpha_{ij}^2)^{1/2}$. (Note that this is in fact the L^2-norm of f as we have already seen in (20).) We can then choose $\tau := \sqrt{2}A$ as a uniform bound on $|\nabla|\nabla f||$. Moreover, as for the constant b from Theorem 1.1, from (21) we get that we can set $b := \sqrt{2}A$ again, i. e. we have $b = \tau$. This leads us to the following result:

Example 8.1. *Let μ be the product measure of n symmetric Bernoulli distributions $\mu_i = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ on $\{\pm 1\}$, and define $f\colon \mathbb{R}^n \to \mathbb{R}$ by*

$$f(x_1, \dots, x_n) := \sum_{i<j}\alpha_{ij}x_ix_j$$

for any real numbers α_{ij}, $i < j$. Set

$$A := (\sum_{i<j}\alpha_{ij}^2)^{1/2}.$$

Then, we have

$$\int \exp\left(\frac{1}{5\sqrt{2}A}|f|\right) d\mu \leq 2.$$

Consequently, we have fluctuations of order 1 if $A = \mathcal{O}(1)$. For instance, if we assume f to be symmetric (that is, invariant under permutations), i. e. $\alpha_{ij} \equiv r$ for all $i < j$ and some $r \in \mathbb{R}$, then this means we need $r = \mathcal{O}(n^{-1})$.

It is possible to replace the constant $5\sqrt{2} \approx 7.1$ in the denominator of the bound given in Example 8.1 by about 6.4. This follows from the proof of Theorem 1.1 if we skip the last few estimates. In particular, we see that we do not lose much in the process of deducing our final inequalities.

We next extend our results to arbitrary functions of n independent Bernoulli variables, i. e. we now allow Hoeffding decompositions which consist of terms from order 2 up to order n. For this, we go back to the deduction of Example 8.1 and see that we can reformulate (23) as

$$|\nabla|\nabla f(x)|| \leq \|f^{\hat{\prime\prime}}(x)\|_{\mathrm{HS}}.$$

Moreover, a close analysis of the deduction of (23) shows that we do not need to assume the Hoeffding term we consider to be of order 2. To see this, take the example of a function f whose Hoeffding decomposition consists of a single term of order d. Then, arguing as in case of $d = 2$ we have

$$(\nabla_i f(x))^2 = (D_i f(x))^2 = \Big(\sum_{\substack{i_1 < \ldots < i_d \\ i \in \{i_1, \ldots, i_d\}}} \alpha_{i_1 \ldots i_d} x_{i_1} \cdots x_{i_d} \Big)^2,$$

so that for $i \neq k$ it follows that

$$(\nabla_i f(x))^2 = \Big(\sum_{\substack{i_1 < \ldots < i_d \\ i \in \{i_1, \ldots, i_d\}}} {}^{(k)}\alpha_{i_1 \ldots i_d} x_{i_1} \cdots x_{i_d} + \sum_{\substack{i_1 < \ldots < i_d \\ i,k \in \{i_1, \ldots, i_d\}}} \alpha_{i_1 \ldots i_d} x_{i_1} \cdots x_{i_d} \Big)^2.$$

Thus, we have

$$(\nabla_i f(x))^2 = (g_{i,k}(x) + D_{ik} f(x))^2,$$

where g_{ik} does not depend on x_k, while we have $D_{ik} f(\sigma_k x) = -D_{ik} f(x)$. In particular, we get

$$(\nabla_i f(\sigma_k(x)))^2 = (g_{i,k}(x) - D_{ik} f(x))^2,$$

while for $i = k$ we have $(\nabla_k f(x))^2 = (\nabla_k f(\sigma_k x))^2$.

This explains why everything we did in case of $d = 2$ also works for $d \geq 2$ arbitrary. Moreover, it is clear that we can apply all these arguments to sums of such terms (possibly from order 2 up to order n) as well, since $(\nabla_i f(x))^2$ will have the same structure again.

Setting $B := \sup_{x \in \{\pm 1\}^n} \|f^{\hat{\prime\prime}}(x)\|_{\mathrm{HS}}$, $\tau := B$ and $b := B$ in Remark 1.3, we therefore arrive at the following generalization of Example 8.1:

Example 8.2. *Let μ be the product measure of n symmetric Bernoulli distributions $\mu_i = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ on $\{\pm 1\}$, and define $f\colon \mathbb{R}^n \to \mathbb{R}$ by*

$$f(x_1, \ldots, x_n) := \sum_{i<j} \alpha_{ij} x_i x_j + \sum_{i<j<k} \alpha_{ijk} x_i x_j x_k + \ldots,$$

where the sum goes up to order n and the $\alpha_{i_1\dots i_d}$ are any real numbers. Set

$$B := \sup_{x\in\{\pm 1\}^n} \|f^{\hat{\prime\prime}}(x)\|_{\mathrm{HS}}.$$

Then, we have

$$\int \exp\left(\frac{1}{5B}|f|\right) d\mu \leq 2.$$

It is possible to sharpen this estimate if we do not set $b = B$ but take the integral $\int \|f^{\hat{\prime\prime}}(x)\|_{\mathrm{HS}} d\mu$. However, since we have $\tau = B$ anyway, this will not change the order of the fluctuations which solely depends on B either way.

Let us check which sort of results we can expect by applying Example 8.2. For this, we again consider a single Hoeffding term but now of order $d > 2$. To start, we take $d = 3$ and assume the Hoeffding term to be symmetric (in order to get a simple result), i. e. we set

$$f(x_1, \dots, x_n) := r \sum_{i<j<k} x_i x_j x_k$$

for some r which we assume to be positive.

Due to (6) we have

$$D_{ij} f(x) = r \sum_{k=1}^{n}{}^{(i,j)} x_i x_j x_k,$$

from which we get

$$B := \sup_{x\in\{\pm 1\}^n} \|f^{\hat{\prime\prime}}(x)\|_{\mathrm{HS}} = r(n-2)\sqrt{n(n-1)}.$$

It therefore follows that

$$\int \exp\left(\frac{1}{5r(n-2)\sqrt{n(n-1)}}|f|\right) d\mu \leq 2,$$

in other words we have fluctuations of order 1 if $r = \mathcal{O}(n^{-2})$.

However, the optimal result would be $r = \mathcal{O}(n^{-3/2})$). In fact, even if we do not apply Example 8.2 but evaluate the behavior of $|\nabla|\nabla f||$ directly we will not arrive at the optimal rate. This is because we are dealing with concentration of second order. In the same way, if we applied a first order concentration result like Proposition 2.1, we would already get a non-optimal result if we considered a second order Hoeffding term as in Example 8.1. In case of a third order Hoeffding term our result would be still worse (in fact, we would get $r = \mathcal{O}(n^{-5/2})$).

Similarly, if we consider a symmetric Hoeffding term of order d for an arbitrary $d \in \{2, 3, \dots, n\}$, i. e. a function f given by

$$f(x_1, \dots, x_n) := r \sum_{i_1<\dots<i_d} x_{i_1} \cdots x_{i_d}$$

for some $r > 0$, we get that

$$\sup_{x \in \{\pm 1\}^n} \|f^{\hat{''}}(x)\|_{\mathrm{HS}} = r\sqrt{n(n-1)}\binom{n-2}{d-2}.$$

Hence, we obtain fluctuations of order 1 if $r = \mathcal{O}(1/(\sqrt{n(n-1)}\binom{n-2}{d-2}))$ or, to put it differently, if $r = \mathcal{O}(n^{-(d-1)})$ in case we regard d as a fixed number independent of n.

It is now possible to combine these results and consider functions whose Hoeffding decompositions consist of several terms. Note that at this point we may even allow first order Hoeffding terms with the help of the results from Section 7 (especially its concluding remarks). In a similar way, we could also discuss general (non-symmetric) functions of independent symmetric Bernoulli variables.

To sum up, we see that our second order results work best for functions whose Hoeffding decompositions are dominated by their second order term. That means, the suprema of the other terms should not be of larger order than the supremum of the second order term. To achieve optimal results for higher order terms we would need higher order analogues of Theorems 1.1 and 1.2.

We conclude this section by comparing some of our results to related work. One example of exponential inequalities for Rademacher chaoses can be found in de la Peña and Giné [D-G], Corollary 3.2.6. We now give a reformulated version of this result:

Proposition 8.3. *Let μ be the product measure of n symmetric Bernoulli distributions $\mu_i = \frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ on $\{\pm 1\}$, and define $f \colon \mathbb{R}^n \to \mathbb{R}$ by*

$$f(x_1, \ldots, x_n) := \alpha_0 + \sum_{i=1}^n \alpha_i x_i + \sum_{i<j} \alpha_{ij} x_i x_j + \sum_{i<j<k} \alpha_{ijk} x_i x_j x_k + \ldots,$$

where the sum goes up to order d for some $d \leq n$ and the $\alpha_{i_1 \ldots i_k}$ are any real numbers. Then, there exists some $\lambda = \lambda(d) \in (0, \infty)$ such that

$$\int \exp\left[\left(\frac{|f|}{\lambda \|f\|_2}\right)^{2/d}\right] d\mu \leq 2.$$

Here, $\|f\|_2$ denotes the L^2-norm of f with respect to μ.

If we once again take

$$f(x_1, \ldots, x_n) := \sum_{i<j} \alpha_{ij} x_i x_j$$

and remember (20), we see that applying Example 8.1 leads to the same result as Proposition 8.3 (up to constants). We cannot expect to get finer estimates as e. g.

Theorem 1.2 in Talagrand [T2], which also provides information about Gaussian type tail probabilities for small t.

On the other hand, let us apply Proposition 8.3 to single Hoeffding terms of some fixed higher order d which does not depend on n. The concentration rates we obtain then depend on the L^2-norm $\|f\|_2$ instead of $\sup_{x\in\{\pm 1\}^n}\|\hat{f''}(x)\|_{\mathrm{HS}}$ as in Example 8.2. In general, this will lead to better estimates than in our case. For instance, Proposition 8.3 yields the optimal rates of concentration for symmetric Hoeffding terms which we could not recover from Example 8.2 in the above discussion.

Finally, consider functions with Hoeffding decompositions with terms up to order n (or some order that grows with n). In this case, Proposition 8.3 will yield a constant λ which depends on n, while our results as stated in Example 8.2 might still allow us to control the higher order terms in a convenient way.

9 Multilinear Polynomials in Independent Random Variables

In this section, we transfer the results from previous section to a more general situation. That is, we consider functions of the same form as in (19), i. e.

$$f(X_1,\dots,X_n) := \alpha_0 + \sum_{i=1}^n \alpha_i X_i + \sum_{i<j} \alpha_{ij} X_i X_j + \dots, \tag{24}$$

where the coefficients α_I (with a suitable multi-index I) are real numbers and the $X_1,\dots,X_n$ are some independent random variables.

Functions of this type are called *multilinear polynomials* in the X_i. Multilinear polynomials appear in many probability theory related fields like random graphs or Boolean functions. Recently, an invariance principle for multilinear polynomials was proved by E. Mossel, R. O'Donnell and K. Oleszkiewicz [M-O-O]. With special regard to concentration inequalities for such functions we also mention the work of W. Schudy and M. Sviridenko [S-S1], [S-S2].

So take n independent random variables $X_1,\dots,X_n$. In order to be able to proceed similarly to the previous section, we assume $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$ for all i. Due to the conditions from Theorem 1.1 we must then require the sum (24) to start with the terms of order 2. Moreover, we have to assume the X_i to be a. s. bounded by some universal constant $M > 0$.

As in the symmetric Bernoulli case, we first consider functions of the form

$$f(X_1,\dots,X_n) := \sum_{i<j} \alpha_{ij} X_i X_j$$

for any real numbers α_{ij}, $i < j$. In particular, it follows from the above discussion that this is a function whose Hoeffding decomposition just consists of a single term of order 2.

The first steps are then exactly the same as in case of the symmetric Bernoulli distribution. The differences only begin with the condition $|\nabla|\nabla f(X_1, \ldots X_n)|| \le 1$ a. s. Instead of (22), we must now work with the more general expression

$$|\nabla|\nabla f(X)|| = \Big(\sum_{k=1}^{n} \frac{1}{2}\bar{\mathbb{E}}_k(|\nabla f(X)| - |T_k \nabla f(X)|)^2\Big)^{1/2} \tag{25}$$

with $X = (X_1, \ldots, X_n)$ and T_k as in Remark 3.3.3.

First, we consider the difference of the two norms in the expectation of (25). Using the assumptions on the moments of the X_i and Remark 3.3.3, we get

$$\begin{aligned}(\nabla_i f(X_1, \ldots, X_n))^2 &= \frac{1}{2}\bar{\mathbb{E}}_i(\sum_{j=1}^{n}{}^{(i)} \alpha_{ij}(X_i - \bar{X}_i)X_j)^2 \\ &= \frac{1}{2}(\sum_{j=1}^{n}{}^{(i)} \alpha_{ij} X_j)^2 (X_i^2 + 1). \end{aligned} \tag{26}$$

In comparison to the symmetric Bernoulli case, we get an additional factor $(X_i^2 + 1)/2$.

Now, as a consequence from (26), for $i = 1, \ldots, n$, $i \neq k$ we have

$$(\nabla_i f(X_1, \ldots, X_n))^2 = \frac{1}{2}(\sum_{j=1}^{n}{}^{(i,k)} \alpha_{ij} X_j + \alpha_{ik} X_k)^2 (X_i^2 + 1)$$

as well as

$$(T_k \nabla_i f(X_1, \ldots, X_n))^2 = \frac{1}{2}(\sum_{j=1}^{n}{}^{(i,k)} \alpha_{ij} X_j + \alpha_{ik} \bar{X}_k)^2 (X_i^2 + 1),$$

while for $i = k$ we have

$$(\nabla_k f(X_1, \ldots, X_n))^2 = \frac{1}{2}(\sum_{j=1}^{n}{}^{(k)} \alpha_{kj} X_j)^2 (X_k^2 + 1)$$

and

$$(T_k \nabla_k f(X_1, \ldots, X_n))^2 = \frac{1}{2}(\sum_{j=1}^{n}{}^{(k)} \alpha_{kj} X_j)^2 (\bar{X}_k^2 + 1).$$

In particular, we do not have $(\nabla_k f)^2 = (T_k \nabla_k f)^2$ anymore. Therefore, these terms will need different treatment than those in the symmetric Bernoulli case.

Hence, we copy our approach from the previous section for all entries but the k-th one. For that, set $a := \sum_{i=1}^{n}{}^{(k)}(\nabla_i f)^2$, $b := (\nabla_k f)^2$, $c := \sum_{i=1}^{n}{}^{(k)}(T_k \nabla_i f)^2$ and $d := (T_k \nabla_k f)^2$ so that $||\nabla f| - |T_k \nabla f|| = |\sqrt{a+b} - \sqrt{c+d}|$.

The simplest idea for estimating $|\sqrt{a+b}-\sqrt{c+d}|$ is as follows: For simplicity, assume that $a,b,c,d>0$. (This is no loss of generality since if this is not the case everything we will do in the sequel can either easily be adapted or gets trivial.) Then, we have

$$\begin{aligned}|\sqrt{a+b}-\sqrt{c+d}| &= \frac{|a+b-c-d|}{\sqrt{a+b}+\sqrt{c+d}}\\ &\leq \frac{|a-c|}{\sqrt{a+b}+\sqrt{c+d}}+\frac{|b-d|}{\sqrt{a+b}+\sqrt{c+d}}\\ &\leq \frac{|a-c|}{\sqrt{a}+\sqrt{c}}+\frac{|b-d|}{\sqrt{b}+\sqrt{d}}\\ &= |\sqrt{a}-\sqrt{c}|+|\sqrt{b}-\sqrt{d}|.\end{aligned}$$

However, this estimate is too weak of be useful. This is because in fact, remembering (25) the second summand will finally yield a contribution of

$$\frac{1}{2}\bar{\mathbb{E}}_k(\nabla_k f - T_k\nabla_k f)^2 = (\nabla_{kk}f)^2.$$

Unlike in case of the difference operator D, we do not have $\nabla_{kk}=\nabla_k$. However, using (26), $\mathbb{E}X_i^2=1$ and $\mathbb{E}\sqrt{X_i^2+1}\leq\sqrt{2}$ (due to Jensen's inequality), we still get

$$\begin{aligned}(\nabla_{kk}f)^2 &= \frac{1}{2}\bar{\mathbb{E}}_k(\nabla_k f - T_k\nabla_k f)^2\\ &= \frac{1}{4}(\sum_{j=1}^n {}^{(k)}\alpha_{kj}X_j)^2\bar{\mathbb{E}}_k(\sqrt{X_k^2+1}-\sqrt{\bar{X}_k^2+1})^2\\ &\geq \frac{1}{4}(\sum_{j=1}^n {}^{(k)}\alpha_{kj}X_j)^2(X_k^2+1-2\sqrt{2}\sqrt{X_k^2+1}+2)\\ &= \frac{1}{4}(\sum_{j=1}^n {}^{(k)}\alpha_{kj}X_j)^2(\sqrt{X_k^2+1}-\sqrt{2})^2.\end{aligned}$$

Comparing this to (26), we see that in essence we would have arrived as a purely first order condition again.

Therefore, we slightly modify the upper estimate. That is, we do not compare the k-th entries against each other but against the norm of the "complete" vector. Indeed, by a simple modification of the above procedure we arrive at the estimate

$$|\sqrt{a+b}-\sqrt{c+d}|\leq|\sqrt{a}-\sqrt{c}|+\frac{|b-d|}{\sqrt{a+b}}.$$

We proceed with estimating $|\sqrt{a}-\sqrt{c}|$. Similarly to the symmetric Bernoulli case, define v to be the vector with entries $(\sum_{j=1}^n {}^{(i,k)}\alpha_{ij}X_j)\sqrt{(X_i^2+1)/2}$, w to be the vector with entries $\alpha_{ik}X_k\sqrt{(X_i^2+1)/2}$ and w' the one with entries $\alpha_{ik}\bar{X}_k\sqrt{(X_i^2+1)/2}$

for all $i \neq k$ in each case (i. e. we consider $(n-1)$-dimensional vectors only). It follows that

$$\begin{aligned}|\sqrt{a}-\sqrt{c}| &= ||v+w|-|v+w'|| \\ &\leq |w-w'| = \big(\frac{1}{2}\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k-\bar{X}_k)^2\big)^{1/2}.\end{aligned}$$

Moreover, in case of $|b-d|/\sqrt{a+b}$, we immediately get

$$\frac{|b-d|}{\sqrt{a+b}} = \frac{1}{\sqrt{2}}\frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{kj}X_j)^2|X_k^2-\bar{X}_k^2|}{(\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1))^{1/2}}.$$

Using these estimates, we take squares and the integral $\bar{\mathbb{E}}_k$. Temporarily omitting the factor $1/2$, this yields

$$\begin{aligned}&\bar{\mathbb{E}}_k\Big(\big(\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k-\bar{X}_k)^2\big)^{1/2} + \frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{ij}X_j)^2|X_k^2-\bar{X}_k^2|}{(\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1))^{1/2}}\Big)^2 \\ =&\bar{\mathbb{E}}_k\Big(\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k-\bar{X}_k)^2 + \frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{ij}X_j)^4(X_k^2-\bar{X}_k^2)^2}{\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1)} \\ &+ 2\big(\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k-\bar{X}_k)^2\big)^{1/2}\frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{ij}X_j)^2|X_k^2-\bar{X}_k^2|}{(\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1))^{1/2}}\Big) \\ \leq&\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k^2+1) + \frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{ij}X_j)^4(X_k^2+M_k^2)^2}{\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1)} \\ &+ 2\big(\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k^2+1)\big)^{1/2}\frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{ij}X_j)^2(X_k^2+M_k^2)}{(\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1))^{1/2}}.\end{aligned}$$

Here, we set $M_k := \operatorname{ess\,sup}|X_k|$. Moreover, we have used Hölder's inequality in the form of

$$\bar{\mathbb{E}}_k|X_k-\bar{X}_k| \leq (\bar{\mathbb{E}}_k(X_k-\bar{X}_k)^2)^{1/2}.$$

Now we define two more vectors $u, u' \in \mathbb{R}^n$ via

$$u_k := \big(\sum_{i=1}^{n} {}^{(k)}\alpha_{ik}^2(X_i^2+1)(X_k^2+1)\big)^{1/2}$$

and

$$u_k' := \frac{(\sum_{j=1}^{n} {}^{(k)}\alpha_{ij}X_j)^2(X_k^2+M_k^2)}{(\sum_{i=1}^{n}(\sum_{j=1}^{n} {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1))^{1/2}}$$

for each $k = 1, \ldots, n$. Summarizing the above discussion and putting everything together (in particular, remember there is second factor $1/\sqrt{2}$ coming from (25)) then reveals that we have shown $|\nabla|\nabla f(x)|| \leq \frac{1}{2}|u+u'|$, which we now simply estimate by $\frac{1}{2}(|u|+|u'|)$.

We can then identify $|u|/2$ as

$$\frac{1}{2}\Big(\sum_{i\neq k}\alpha_{ik}^2(X_i^2+1)(X_k^2+1)\Big)^{1/2}, \tag{27}$$

and in case of $|u'|/2$ we have

$$\begin{aligned}&\frac{1}{2}\left(\frac{\sum_{i=1}^n(\sum_{j=1}^n {}^{(i)}\alpha_{ij}X_j)^4(X_i^2+M_i^2)^2}{\sum_{i=1}^n(\sum_{j=1}^n {}^{(i)}\alpha_{ij}X_j)^2(X_i^2+1)}\right)^{1/2}\\ &\leq \frac{1}{\sqrt{2}}M\max_{i=1,\ldots,n}\sqrt{M_i^2+1}|\sum_{j=1}^n {}^{(i)}\alpha_{ij}X_j|. \end{aligned}\tag{28}$$

In the latter inequality we have used the simple estimate

$$\sum_i x_i^2y_i^2 \leq \sum_i x_iz_i(y_i/z_i)\max_i x_iy_i$$

for positive x_i, y_i, z_i. Moreover, remember that $M \geq \max_i M_i$ is an upper bound on the X_i.

It is now possible to relate (27) and (28) to the difference operators from Section 3 again. As for (27), note that we have

$$\|f^{\hat{''}}(X_1,\ldots,X_n)\|_{\mathrm{HS}} = \Big(\sum_{i\neq j}\alpha_{ij}^2X_i^2X_j^2\Big)^{1/2}.$$

Therefore, we can interpret (27) as the value of the Hilbert-Schmidt norm of the Hessian $f^{\hat{''}}(X_1,\ldots,X_n)$ if in each of the components $f_{ij}^{\hat{''}}(X_1,\ldots,X_n)$ we replace the random variables X_i and X_j by $\sqrt{(X_i^2+1)/2}$ and $\sqrt{(X_j^2+1)/2}$, respectively.

Moreover, using (26) we see that in fact, (28) is nothing but

$$M\max_{i=1,\ldots,n}\operatorname{ess\,sup}{}_{(i)}|\nabla_i f(X_1,\ldots,X_n)|.$$

Here, $e\mathring{s}s\sup_{(i)}$ means that we only take the essential supremum in X_i but not (yet) in the other random variables. The appearance of first order differences is a consequence of the structures of $\nabla_k f$ and $T_k\nabla_k f$ which we have discussed after deducing (26).

As in the symmetric Bernoulli case, we now note that all we did so far actually does not depend on the fact that we have dealt with a second order Hoeffding term only, but we can choose any multilinear polynomial with terms from order 2 up to n instead. Again, the background is that for $i \neq k$ we basically have

$$(\nabla_i f(X_1,\ldots,X_n))^2 = (g_{ik}(X_1,\ldots,X_n) + g'_{ik}(X_1,\ldots,X_n))^2$$

for a function g_{ik} which does not depend on X_k and a function g'_{ik} which is closely related to $D_{ik}f$.

Finally, in some applications it is more convenient to allow arbitrary independent random variables and then to consider suitable compositions $\psi_i(X_i)$. So let $X_1, \dots, X_n$ be independent random variables with distributions $\mu_1, \dots, \mu_n$ and product measure $\mu = \otimes_{i=1}^n \mu_i$ together with measurable functions $\psi_i \colon \mathbb{R} \to \mathbb{R}$ such that we have

$$\mathbb{E}\psi_i(X_i) = 0, \qquad \mathbb{E}\psi_i(X_i)^2 = 1 \qquad \text{and} \qquad \max_i \sup_{x \in \mathrm{supp}(\mu_i)} |\psi_i(x)| \leq M. \tag{29}$$

Here, M is some universal constant, and $\mathrm{supp}(\mu_i)$ denotes the support of μ_i.

The functions to consider are then of the form

$$f(x_1, \dots, x_n) := \sum_{i<j} \alpha_{ij}\psi_i(x_i)\psi_j(x_j) + \sum_{i<j<k} \alpha_{ijk}\psi_i(x_i)\psi_j(x_j)\psi_k(x_k) + \dots, \tag{30}$$

where the sum goes up to order n and the $\alpha_{i_1 \dots i_d}$ are real numbers.

The modified Hessian we introduced in the context of (27) can now be defined as follows:

Definition 9.1. *Let f be a function as in* (30)*, and let $f^{\hat{\prime\prime}}(x)$ be its dediagonalized Hessian with respect to the difference operator D. Then, we denote by $f^{\hat{\prime\prime}*}(x)$ the $n \times n$-matrix which we get if for all $i \neq j$, in the ij-th entry of $f^{\hat{\prime\prime}}(x)$ we replace the functions $\psi_i(x_i)$ and $\psi_j(x_j)$ by $\sqrt{(\psi_i(x_i)^2 + 1)/2}$ and $\sqrt{(\psi_j(x_j)^2 + 1)/2}$, respectively.*

Based on this, we choose real numbers B_1 and B_2 such that

$$B_1 \geq \sup_{x \in \mathrm{supp}(\mu)} \|f^{\hat{\prime\prime}*}(x)\|_{\mathrm{HS}} \tag{31}$$

and moreover

$$B_2 \geq M \sup_{x \in \mathrm{supp}(\mu)} \max_{i=1,\dots,n} |\nabla_i f(x)|, \tag{32}$$

where $\mathrm{supp}(\mu)$ denotes the support of μ.

We then arrive at the following result:

Example 9.2. *Consider n independent random variables $X_1, \dots, X_n$ with distributions μ_i, and let $\mu = \otimes_{i=1}^n \mu_i$ be their product measure. Furthermore, let $\psi_i \colon \mathbb{R} \to \mathbb{R}$ be measurable functions as in* (29)*. Consider the function $f \colon \mathbb{R}^n \to \mathbb{R}$ which is given by*

$$f(x_1, \dots, x_n) := \sum_{i<j} \alpha_{ij}\psi_i(x_i)\psi_j(x_j) + \sum_{i<j<k} \alpha_{ijk}\psi_i(x_i)\psi_j(x_j)\psi_k(x_k) + \dots,$$

where the sum goes up to order n and the $\alpha_{i_1 \dots i_d}$ are real numbers. Choose B_1 and B_2 as in (31) *and* (32)*. Then, we have*

$$\int \exp\left(\frac{1}{8(B_1 + B_2)}|f|\right) d\mu \leq 2.$$

Proof. Most of the proof directly follows from the above discussion. In particular, we apply Remark 1.3 with $\tau = B_1 + B_2$. It only remains to check that it is possible also to set $b = B_1 + B_2$.

For this, note that by definition we have

$$b^2 \geq \mathbb{E}\|f^{\widehat{\prime\prime}}(X_1, \ldots, X_n)\|_{\mathrm{HS}}^2.$$

The entries of $f^{\widehat{\prime\prime}}(X_1, \ldots, X_n)$ are of the form

$$f_{ij}^{\widehat{\prime\prime}}(X_1, \ldots, X_n) = D_{ij} f(X_1, \ldots, X_n) = \psi_i(X_i)\psi_j(X_j) g_{ij}(X_1, \ldots, X_n)$$

for some function g_{ij} which does not depend on X_i and X_j. Moreover, we have

$$f_{ij}^{\widehat{\prime\prime}*}(X_1, \ldots, X_n) = \sqrt{\frac{\psi_i(X_i)^2 + 1}{2}} \sqrt{\frac{\psi_j(X_j)^2 + 1}{2}} g_{ij}(X_1, \ldots, X_n)$$

for the same function g_{ij}.

Since $\mathbb{E}\psi_i(X_i)^2 = 1$ for all $i = 1, \ldots, n$, it follows that

$$\mathbb{E} f^{\widehat{\prime\prime}}(X_1, \ldots, X_n)_{ij}^2 = \mathbb{E} f^{\widehat{\prime\prime}*}(X_1, \ldots, X_n)_{ij}^2$$

for any $i \neq j$. Using (31), we therefore get that

$$B_1^2 \geq \mathbb{E}\|f^{\widehat{\prime\prime}*}(X_1, \ldots, X_n)\|_{\mathrm{HS}}^2 = \mathbb{E}\|f^{\widehat{\prime\prime}}(X_1, \ldots, X_n)\|_{\mathrm{HS}}^2,$$

which completes the proof. □

To sum up, Example 9.2 enables us to replace the condition $|\nabla|\nabla f|| \leq 1$ by conditions solely depending on the dediagonalized Hessian of f (or simple modifications of it) and first order difference operators.

As the case of a single Hoeffding term of second order is of particular importance, we also give a reformulation of Example 9.2 for this situation. Setting

$$A_1 := (\sum_{i<j} \alpha_{ij}^2)^{1/2} \qquad \text{and} \qquad A_2 := \max_{i=1,\ldots,n} \sum_{j=1}^{n}{}^{(i)} |\alpha_{ij}|,$$

it is possible to take $B_1 := (M^2 + 1)A_1/2$ and $B_2 := M^2\sqrt{(M^2 + 1)/2}A_2$.

Therefore, applying Remark 1.3 with $\tau := B_1 + B_2$ and $b := \sqrt{2}A_1$ leads us to the following analogue of Example 8.1:

Example 9.3. *Let $X_1, \ldots, X_n$ be independent random variables with distributions μ_i, and let $\mu := \otimes_{i=1}^n \mu_i$ be the product measure. Consider the function $f \colon \mathbb{R}^n \to \mathbb{R}$ which is given by*

$$f(x_1, \ldots, x_n) := \sum_{i<j} \alpha_{ij} \psi_i(x_i) \psi_j(x_j)$$

with functions ψ_i as described in (29) *and real numbers α_{ij}. Define A_1, A_2, B_1 and B_2 as above. Then, we have*

$$\int \exp\left(\frac{1}{6(B_1+B_2)+4A_1^2/(B_1+B_2)}|f|\right)d\mu \le 2.$$

As a simple application, we continue the discussion of Bernoulli variables which we started in the previous section. The aim is getting an analogue of Example 8.1 for the non-symmetric case. So, consider the product measure μ of n Bernoulli measures $\mu_i := p\delta_{+1} + (1-p)\delta_{-1}$ for some $p \in (0,1)$.

As the expected value of the measures μ_i is $2p-1$ and the variance is $4p(1-p)$, we consider

$$\begin{aligned} f(x_1,\dots,x_n) &:= \sum_{i<j} \alpha_{ij}(x_i-(2p-1))(x_j-(2p-1)) \\ &= 4p(1-p)\sum_{i<j}\alpha_{ij}\psi(x_i)\psi(x_j) := 4p(1-p)\tilde{f}(x_1,\dots,x_n) \end{aligned}$$

Here, we set

$$\psi(x) = (x-(2p-1))/\sqrt{4p(1-p)},$$

so that we have

$$|\psi_i(x)| \le 2\max(p,1-p)/\sqrt{4p(1-p)}$$

on the relevant domain. If we assume $p \in (0,1/2]$ we can therefore take $M := \sqrt{(1-p)/p}$ as an upper bound on $|\psi|$.

Now define A_1, A_2, B_1 and B_2 as above with respect to $\tilde{f}$. Then, applying Example 9.3 (where we can replace the factor 6 in the denominator by 3 as we are in a two-point situation) leads us to the inequality

$$\int \exp\left(\frac{1}{(3(B_1+B_2)+4A_1^2/(B_1+B_2))4p(1-p)}|f|\right)d\mu \le 2.$$

Consequently, if A_2 is of the same order as A_1 (or a smaller one), we get the same behavior as in the symmetric case but just with an additional dependency on p.

Let us discuss the differences between Examples 8.1 and 8.2 and Examples 9.2 and 9.3 in general. We particularly focus on the additional term A_2 (or B_2, in general). We start with functions of the form

$$f(x_1,\dots,x_n) := \sum_{i<j} r\psi_i(x_i)\psi_j(x_j)$$

for some real number r which we assume to be positive. Then, we immediately get

$$A_1 = r\sqrt{n(n-1)} \qquad \text{and} \qquad A_2 = r(n-1).$$

As we have $A_2 \le A_1$, we will replace A_2 by A_1 in the sequel.

As in the symmetric Bernoulli case, we therefore have fluctuations of order 1 if we have $r = \mathcal{O}(n^{-1})$. The only difference is that we now also have a dependency on the bound M. To illustrate this, we use the rough estimate

$$B_1 + B_2 \leq \Big(\frac{M^2+1}{2} + M^3\Big) A_1 \leq 2M^3 A_1$$

as we have $M \geq 1$ (and consequently $\sqrt{(M^2+1)/2} \leq M$, in particular). Setting $\tau = 2M^3 A_1$ in the deduction of Example 9.3, we get

$$\int \exp\left(\frac{1}{(12M^3 + 2M^{-3}) r \sqrt{n(n-1)}} |f|\right) d\mu \leq 2.$$

We can proceed in the same way for any symmetric Hoeffding term of order $d \geq 2$ and finally for any symmetric function whose Hoeffding decomposition starts with terms of at least second order. In doing so, we get the same results as in case of symmetric functions of symmetric Bernoulli variables but with an additional dependency on the upper bound M. This also holds for non-symmetric functions if A_2 is not of larger order than A_1.

So, it remains to check the situations in which the order of A_2 is larger than the order of A_1. We consider a single Hoeffding term of order 2 again. Then, the cases to consider are those in which there is a small number of rows (or, equivalently, columns) $(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in})$ which "dominate" the matrix consisting of the α_{ij}.

For instance, for $i \neq j$, set $\alpha_{ij} = r$ for some $r > 0$ if $i = 1$ or $j = 1$ and 0 if not, i. e. we consider

$$f(X_1, \ldots, X_n) := r\psi_1(X_1) \sum_{j=2}^{n} \psi_j(X_j). \tag{33}$$

We obviously have $A_1 = r\sqrt{n-1}$ and $A_2 = r(n-1)$. So by applying Example 9.3, we see that we get fluctuations of order 1 if $r = \mathcal{O}(n^{-1})$.

On the other hand, we clearly have

$$|f(X_1, \ldots, X_n)| \leq rM |\sum_{i=2}^{n} \psi_i(X_i)|,$$

which means f is dominated by a statistic of order 1. Hence we should already expect fluctuations of order 1 if $r = \mathcal{O}(n^{-1/2})$. Similar to the case of the third or higher order Hoeffding terms we studied at the end of Section 8, we therefore get non-optimal results at this point because we are not in a proper second order situation.

Indeed, the situation of A_2 dominating A_1 will occur if we are dealing with a "degenerated" second order statistic for which there is a first order term dominating the behavior of f. Calculating $|\nabla|\nabla f||$ directly in simple examples confirms these observations. In particular, we see that unlike in the symmetric Bernoulli case we shall need A_2 (or B_2 in general) as well.

To conclude this section, we compare our results to related work again. If we focus on second order Hoeffding terms, a possible object of comparison is the Hanson-Wright inequality. To state it, we quote M. Rudelson and R. Vershynin [R-V], Theorem 1.1:

Theorem 9.4 (Hanson-Wright inequality). *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfy*

$$\mathbb{E}X_i = 0 \qquad \textit{and} \qquad \|X_i\|_{\psi_2} := \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p} \leq K.$$

Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,

$$\mathbb{P}\left(|X^{\mathrm{T}}AX - \mathbb{E}X^{\mathrm{T}}AX| > t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^4 \|A\|_{\mathrm{HS}}}, \frac{t}{K^2 \|A\|_{2 \to 2}}\right)\right).$$

Here, $\|A\|_{2\to 2} := \sup_{x \neq 0} \|Ax\|_2 / \|x\|_2$ denotes the operator norm of A with respect to the Euclidean norm on $\mathbb{R}^n$, and c is some positive absolute constant.

Let us compare this to Example 9.3. For that, take an upper triangular matrix $A = (\alpha_{ij})_{i<j}$ and n independent random variables $X_1, \ldots, X_n$ which we assume to be centered, to have unit variance and to be bounded by some constant M. This is necessary for applying Example 9.3, and of course it also implies that the subgaussian tails condition $\|X_i\|_{\psi_2} \leq K$ is fulfilled.

We then have a function

$$f(X_1, \ldots, X_n) = \sum_{i<j} \alpha_{ij} X_i X_j$$

whose Hoeffding decomposition consists of the second order term only indeed. Note that we can rewrite the quantities A_1 and A_2 from Example 9.3 as

$$A_1 = \|A\|_{\mathrm{HS}} \qquad \text{and} \qquad A_2 = \|A\|_{\infty \to \infty},$$

where $\|A\|_{\infty\to\infty} := \sup_{x \neq 0} \|Ax\|_\infty / \|x\|_\infty$ denotes the operator norm of A with respect to the supremum norm $\|x\|_\infty := \max_i |x_i|$ on $\mathbb{R}^n$. (In principle, by a suitable adaption of Definition 9.1, we could even rewrite B_2 from (32) this way.)

Then, applying Example 9.3 and Chebychev's inequality leads us to

$$\mathbb{P}\left(|f| \geq t\right) \leq 2 \exp\left(-Ct\right)$$

for any $t > 0$. Here, $C = C(M, \|A\|_{\mathrm{HS}}, \|A\|_{\infty\to\infty})$ is some absolute constant. Possible values for C are given in Example 9.3.

To sum up, we cannot recover the Gaussian rate for small t from Theorem 9.4, but the non-Gaussian part of the rate function in the Hanson-Wright inequality only differs from ours by the matrix norms involved. Recalling our discussion in the context of (33), in typical second order situations we can therefore expect to arrive at the same rates up to constants. This extends to more general cases like Hoeffding terms of higher order if we compare results like Theorem 4.1.12 in de la Peña and Giné [D-G] or Theorem 8.3 in Major [M] to Example 9.2.

Finally, since we are considering polynomials we can also refer to Theorem 1.4 in Adamczak and Wolff [A-W]. Some annotations on those results as compared to ours have already been made in Section 2, and we will not go deeper into the details here.

10 Second Order Concentration of Empirical Distribution Functions

The following section deals with concentration of empirical distribution functions. In doing so, we lean on S. G. Bobkov and F. Götze [B-G3] (particularly Sections 6 and 7). Our aim is to transfer some of their results to the second order situation.

The general situation is as follows: We consider a family of independent random variables ξ_j with distributions $\mathbb{P}_j$, $j = 1, \dots N$, where N is any natural number. Let $\mathbb{P} = \otimes_{j=1}^N \mathbb{P}_j$ be their product measure. On this probability space $(\mathbb{R}^N, \mathbb{P})$, we consider a random vector $(X_1, \dots, X_n)$ in $\mathbb{R}^n$ with some distribution μ. Here, n is some natural number which might or might not coincide with N.

For instance, we might take $N = n$ and $\xi_i = X_i$ for all i. Another possible situation is switching to double indexes ξ_{jk}, $1 \leq j \leq k \leq n$ for some $n \in \mathbb{N}$, setting $\xi_{kj} \equiv \xi_{jk}$ for all $j \neq k$ and considering the symmetric random matrix W which is defined by $W_{jk} := \xi_{jk}/\sqrt{n}$. (We thus have $N = n(n-1)/2$.) Then, a random vector of particular interest is the collection of the eigenvalues $(X_1, \dots, X_n)$ of W, where we assume $X_1 \leq \dots \leq X_n$.

In any such situation, we want to study the fluctuations of the empirical distribution function

$$F_n(y) := \frac{1}{n} \mathrm{card}\{i \leq n \colon X_i \leq y\}, \qquad y \in \mathbb{R}.$$

Its expected value with respect to $\mathbb{P}$ is the mean empirical distribution function

$$F(y) := \mathbb{E} F_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i \leq y).$$

Moreover, with respect to the random variables ξ_j, for any fixed $y \in \mathbb{R}$, $F_n(y)$ has Hoeffding decomposition

$$F_n(y) = F_{n,0}(y) + F_{n,1}(y) + F_{n,2}(y) + \dots + F_{n,N}(y).$$

Note that $F_{n,0}(y) \equiv F(y)$ by Theorem 3.4.

In [B-G3], the fluctuations of F_n around its expectation F were considered by studying the Kolmogorov metric, for example. In our situation, we will work with the difference operators we introduced in Section 3 and study the fluctuations of $F_n - F_{n,1} - F$, i. e. we also subtract the first order Hoeffding term.

By Remark 1.3, if for some $y \in \mathbb{R}$ the conditions from Theorem 1.1 with $b = \tau$ (depending on y) are fulfilled, we have

$$\int e^{|F_n(y) - F_{n,1}(y) - F(y)|/(8\tau)} d\mathbb{P} \leq 2.$$

By applying Chebychev's inequality, this immediately yields a result parallel to [B-G3], Proposition 6.3:

Proposition 10.1. *In the situation as described above, fix $y \in \mathbb{R}$ and assume that*

$$|\nabla|\nabla(F_n(y) - F_{n,1}(y) - F(y))|| \leq \tau$$

on the support of $\mathbb{P}$ as well as

$$\int \|F_n(y)^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mathbb{P} \leq \tau^2$$

for some $\tau \geq 0$. Then, for any $r > 0$ we have

$$\mathbb{P}(|F_n(y) - F_{n,1}(y) - F(y)| \geq \tau r) \leq 2e^{-r/8}.$$

In particular, with some absolute constant C, we have

$$\mathbb{E}|F_n(y) - F_{n,1}(y) - F(y)| \leq C\tau.$$

More precisely, it is possible to take $C = 16$.

Unlike in Proposition 6.3 from [B-G3], we do not have to assume that F has a density. The background is that F having a density encodes a kind of first order boundedness condition, which we have replaced by the conditions from Theorem 1.1. Moreover, note that we have used $\|F_n(y)^{\hat{\prime\prime}}\|_{\mathrm{HS}} = \|(F_n(y) - F_{n,1}(y) - F(y))^{\hat{\prime\prime}}\|_{\mathrm{HS}}$ due to (18).

To see how Proposition 10.1 differs from Proposition 6.3 in [B-G3], we consider the simple example of $N = n$ and $\xi_i = X_i$. Then, we can write $F_n(y) = \sum_{i=1}^n \mathbb{1}_{(-\infty,y]}(X_i)$, and it is clear that the Hoeffding decomposition with respect to the X_i consists of terms of order 0 and 1 only. Therefore, we have $F_n - F_{n,1} - F \equiv 0$, and we can thus set $\tau = 0$. Proposition 10.1 then yields that we have zero tails indeed. This is of course a trivial result, but it illustrates that by considering second order concentration the situation can change quite remarkably.

We can also adapt Proposition 6.4 from [B-G3], which gives estimates in the case of replacing the mean value $F(y)$ by some canonical probability distribution function $G(y)$ (for instance the semi-circle distribution in case of eigenvalues of a random matrix). Here, the basic argument for adapting it is as follows: for any $\lambda > 0$, we have

$$\begin{aligned}\int e^{\lambda|F_n(y)-F_{n,1}(y)-G(y)|}d\mathbb{P} &\le \int e^{\lambda(|F_n(y)-F_{n,1}(y)-F(y)|+|F(y)-G(y)|)}d\mathbb{P}\\ &\le e^{\lambda\|F-G\|}\int e^{\lambda|F_n(y)-F_{n,1}(y)-F(y)|}d\mathbb{P},\end{aligned}$$

where $\|\cdot\|$ denotes the Kolmogorov distance, i. e.

$$\|F-G\| = \sup_{y\in\mathbb{R}}|F(y)-G(y)|$$

for any two probability distribution functions F and G on the real line.

Going back to the proof of Theorem 1.1 in Section 6, we now take $\lambda = 1/(2\sigma\tilde{\sigma})$ and combine the estimate from above with (17). (In particular, note that at this point we assume $|\nabla|\nabla(F_n(y)-F_{n,1}(y)-F(y))|| \le 1$.) This leads us to

$$\begin{aligned}&\int \exp\left(\frac{1}{2\sigma\tilde{\sigma}}\left(|F_n(y)-F_{n,1}(y)-F(y)|-\|F-G\|\right)\right)d\mathbb{P}\\ &\le 2\exp\left(\frac{1}{2\tilde{\sigma}^2}\frac{1}{d-1}\int\|F_n(y)^{\hat{''}}\|_{\mathrm{HS}}^2 d\mathbb{P}\right).\end{aligned}$$

We can then proceed in the same way as in the rest of the proof of Theorem 1.1 and finally get

$$\int \exp\left(\frac{1}{2(3+b^2/(d-1))}\left(|F_n(y)-F_{n,1}(y)-F(y)|-\|F-G\|\right)\right)d\mathbb{P} \le 2$$

in parallel to Theorem 1.1 (given that the respective second order conditions hold). We can also make similar conclusions as in Remark 1.3, so that we can finally assume $|\nabla|\nabla(F_n(y)-F_{n,1}(y)-F(y))|| \le \tau$ for some $\tau \ge 0$.

As a result, we get the following analogue of Proposition 6.4 from [B-G3]:

Proposition 10.2. *In the situation of Proposition 10.1, fix $y \in \mathbb{R}$ and assume that*

$$|\nabla|\nabla(F_n(y)-F_{n,1}(y)-F(y))|| \le \tau$$

on the support of $\mathbb{P}$ as well as

$$\int\|F_n(y)^{\hat{''}}\|_{\mathrm{HS}}^2 d\mathbb{P} \le \tau^2$$

for some $\tau \ge 0$. Moreover, let G be any distribution function on the real line. Then, for any $r > 0$ we have

$$\mathbb{P}(|F_n(y)-F_{n,1}(y)-G(y)| \ge \tau r + \|F-G\|) \le 2e^{-r/8}.$$

In particular, up to some absolute constant C,

$$\mathbb{E}|F_n(y)-F_{n,1}(y)-G(y)| \le C\tau + \|F-G\|.$$

As in Proposition 10.1, it is possible to take $C = 16$.

In contrast to Proposition 6.4 from [B-G3], we still have to take F instead of G in the second order conditions. The reason is that as our work is based on Hoeffding decompositions, we must take care of removing the terms of order 0 and 1, i. e. in particular the expected value.

We go on with formulating second order analogues of Theorem 1.2 and Theorem 7.1 from [B-G3]. These theorems give estimates for the Kolmogorov distance $\|F_n - F\|$. In the second order setting, we will rather get estimates for

$$\sup_{y \in \mathbb{R}} |F_n(y) - F_{n,1}(y) - F(y)|,$$

which is not a Kolmorogov distance in the proper sense because it is not a difference of two distribution functions due to the additional first order Hoeffding term. In principle, this is no problem, and we will keep denoting it $\|F_n - F_{n,1} - F\|$. However, we cannot assume this quantity to be bounded by 1 anymore, which entails some minor changes in the theorems we state.

Before we go on to the results, we need to adapt inequalities (6.5) and (6.6) from [B-G3]. This can easily be achieved by applying Proposition 10.1, which yields the bounds

$$\mathbb{P}(F_n(y) - F_{n,1}(y) - F(y) \geq \tau r) \leq 2e^{-r/8} \tag{34}$$

as well as

$$\mathbb{P}(F(y) + F_{n,1}(y) - F_n(y) \geq \tau r) \leq 2e^{-r/8}. \tag{35}$$

Now we can formulate a second order version of Theorem 1.2 from [B-G3].

Theorem 10.3. *In the situation of Proposition 10.1, assume that F is continuous and that for any $y \in \mathbb{R}$ we have*

$$|\nabla|\nabla(F_n(y) - F_{n,1}(y) - F(y))|| \leq \tau$$

on the support of $\mathbb{P}$ as well as

$$\int \|F_n(y)^{\hat{''}}\|_{\mathrm{HS}}^2 d\mathbb{P} \leq \tau^2$$

for some $\tau \geq 0$ which does not depend on y. Then, for any $r > 0$ we have

$$\mathbb{P}(\|F_n - F_{n,1} - F\| \geq r) \leq \frac{8}{r} e^{-\gamma r/\tau}.$$

In particular,

$$\mathbb{E}\|F_n - F_{n,1} - F\| \leq C(8\tau + \tau^2) \log\Big(1 + \frac{1}{\tau}\Big).$$

Here, γ and C are positive absolute constants.

Proof. The proof of the first statement works in the same way as the proof of its analogue in Theorem 1.2 from [B-G3]. The only difference is we have to replace (6.5) with (34) and (6.6) with (35). Note that we need to assume F to be continuous since otherwise we would not be able to find points $-\infty = y_0 \leq y_1 \leq \ldots \leq y_{k-1} \leq y_k = +\infty$ for any $k \in \mathbb{N}$ such that

$$F(y_i) - F(y_{i-1}) \leq \frac{1}{k}, \qquad i = 1, \ldots, k,$$

as in (7.1) from [B-G3]. We finally get $\gamma = 1/16$.

The proof of the second statement also works similarly to the one of its analogue in [B-G3], but here we must keep in mind that we cannot assume $\|F_n - F_{n,1} - F\|$ to be bounded by 1 anymore. For any $r_0 > 0$, write

$$\begin{aligned} \mathbb{E}\|F_n - F_{n,1} - F\| &= \int_0^\infty \mathbb{P}(\|F_n - F_{n,1} - F\| \geq r)dr = \int_0^{r_0} + \int_{r_0}^\infty \\ &\leq r_0 + \frac{128\tau}{r_0} \exp(-r_0/(16\tau)), \qquad (*) \end{aligned}$$

where we have used the first part of the theorem.

Now, set $r_0 = 32\tau \log(1 + \frac{1}{\tau})$. Then, the second term in $(*)$ becomes

$$\begin{aligned} \frac{128\tau}{r_0} \exp(-r_0/(16\tau)) &= \frac{4}{\log(1 + 1/\tau)} e^{-2\log(1+1/\tau)} \\ &= \frac{4\tau^2}{(1+\tau)^2 \log(1 + 1/\tau)} \\ &\leq 4B\tau^2 \log\left(1 + \frac{1}{\tau}\right) \end{aligned}$$

with some constant B satisfying $(1+\tau)\log(1+1/\tau) \geq (\frac{1}{B})^{1/2}$. For instance, we can take $B = 1$, and then, by $(*)$, we have

$$\begin{aligned} \mathbb{E}\|F_n - F_{n,1} - F\| &\leq 32\tau \log\left(1 + \frac{1}{\tau}\right) + 4\tau^2 \log\left(1 + \frac{1}{\tau}\right) \\ &\leq 4(8\tau + \tau^2) \log\left(1 + \frac{1}{\tau}\right), \end{aligned}$$

which finishes the proof. In particular, we see that we can take $C = 4$. □

To conclude this section, we give an analogue of Theorem 7.1 from [B-G3], i. e. in our case a version of Theorem 10.3 with F being replaced by some canonical distribution function G again. The proof is similar to the ones of Proposition 10.2 and Theorem 10.3.

Theorem 10.4. *In the situation of Proposition 10.1, assume that for any $y \in \mathbb{R}$ we have*

$$|\nabla|\nabla(F_n(y) - F_{n,1}(y) - F(y))|| \leq \tau$$

on the support of $\mathbb{P}$ as well as

$$\int \|F_n(y)^{\hat{\prime\prime}}\|_{\mathrm{HS}}^2 d\mathbb{P} \leq \tau^2$$

for some $\tau \geq 0$ which does not depend on y. Moreover, let G be any continuous distribution function on the real line. Then, for any $r > 0$ we have

$$\mathbb{P}(\|F_n - F_{n,1} - G\| \geq r + \|F - G\|) \leq \frac{8}{r} e^{-\gamma r/\tau}.$$

In particular,

$$\mathbb{E}\|F_n - F_{n,1} - G\| \leq C(8\tau + \tau^2) \log\Big(1 + \frac{1}{\tau}\Big) + \|F - G\|.$$

Here, γ and C are the same positive absolute constants as in Theorem 10.3.

11 Empirical Distribution Functions: The Bernoulli Case

In this section, we apply some of the results about second order concentration of empirical distribution functions from the previous section. It would be desirable to combine them with the rules of calculus for second order concentration for multilinear polynomials from Sections 8 and 9.

In general, however, the Hoeffding decomposition of an empirical distribution function $F_n(y) := \frac{1}{n}\sum_i \mathbb{1}_{\{X_i \leq y\}}$ with respect to the random variables $\xi_1, \ldots, \xi_N$ does not have the form of a multilinear polynomial. Yet, there is one exception, namely the case that all the ξ_i have symmetric Bernoulli distributions on $\{\pm 1\}$. In this situation, we can use our results from Section 8.

It is easy to derive some basic concentration properties in the symmetric Bernoulli situation. Assume that we have $X_i = \varphi_i(\xi_1, \ldots, \xi_N)$ for all $i = 1, \ldots, n$, where the φ_i are any functions on $\{\pm 1\}^N$. (We do not have to require any restrictions on these functions at this general stage.)

Now, our aim is to apply Proposition 10.2. By Example 8.2, a possible value for τ as in Proposition 10.2 is given by

$$\tau = \tau(y) = \sup_{x \in \{\pm 1\}^N} \|F_n(y)^{\hat{''}}(x)\|_{\mathrm{HS}}.$$

Moreover, for any $1 \leq k \neq l \leq N$, by using Remark 3.3.1 we get

$$\begin{aligned} \nu_{kl}(x,y) :=& D_{kl}F_n(y)(x) \\ =& \frac{1}{4n}\sum_{i=1}^n \Big(\mathbb{I}(\varphi_i(x) \leq y) - \mathbb{I}(\varphi_i(\sigma_k x) \leq y) \\ & - \mathbb{I}(\varphi_i(\sigma_l x) \leq y) + \mathbb{I}(\varphi_i(\sigma_{kl} x) \leq y)\Big), \end{aligned} \tag{36}$$

where $\mathbb{I}(A)$ denotes the indicator function of some event A.

As a first step, we can therefore formulate a pretty general second order concentration result for $F_n(y)$:

Example 11.1. *Let N, n be natural numbers, and consider N independent symmetric Bernoulli variables $\xi_1, \ldots, \xi_N$ each taking values in $\{\pm 1\}$. Set $X_i := \varphi_i(\xi_1, \ldots, \xi_N)$ for all $i = 1, \ldots, n$, where the φ_i are any functions on $\{\pm 1\}^N$. Let $F_n(y)$ be the empirical distribution function of the random variables X_i, and denote its Hoeffding terms with respect to the ξ_i by $F_{n,d}(y)$, $d = 0, 1, \ldots, N$. With ν_{kl} as defined in* (36) *and any $y \in \mathbb{R}$, set*

$$\tau(y) := \sup_{x \in \{\pm 1\}^N} \Big(\sum_{1 \le k \ne l \le N} \nu_{kl}(x, y)^2 \Big)^{1/2}.$$

Then, for any $r > 0$ we have

$$\mathbb{P}(|F_n(y) - F_{n,1}(y) - F(y)| \ge \tau(y) r) \le 2e^{-r/5}$$

as well as

$$\mathbb{E}|F_n(y) - F_{n,1}(y) - F(y)| \le C\tau(y),$$

where C is some absolute constant.

Here we have replaced the bound given in Proposition 10.2 by its slightly better analogue from Example 8.2.

Next, we must to evaluate the behavior of the ν_{kl}. We start with some simples examples which give a flavor of what we can expect. For instance, let $N = n$ and take

$$X_i := \varphi_i(\xi_1, \ldots, \xi_n) := \begin{cases} \xi_i \xi_{i+1}, & i \in \{1, \ldots, n-1\}, \\ \xi_n \xi_1, & i = n. \end{cases} \tag{37}$$

Here we have $\nu_{kl} \equiv 0$ if $|k - l| > 1$ and $\{k, l\} \ne \{1, n\}$. Otherwise we have expressions of the form

$$\nu_{i,i+1}(x, y) = \frac{1}{2n} \Big(\mathbb{I}(x_i x_{i+1} \le y) - \mathbb{I}(-x_i x_{i+1} \le y) \Big).$$

These terms can only be non-zero if $y \in [-1, +1)$, and it is easily seen that

$$\begin{aligned} \tau(y) &= \sup_{x \in \{\pm 1\}^n} \Big(\sum_{1 \le k \ne l \le n} \nu_{kl}(x, y)^2 \Big)^{1/2} \\ &= \Big(2n \frac{1}{4n^2} \Big)^{1/2} = \frac{1}{\sqrt{2n}} \end{aligned}$$

for any $y \in [-1, +1)$.

Let us briefly check how this changes when we introduce additional coefficients $a_i \in \mathbb{R}$, i. e. we take

$$\varphi_i(\xi_1, \ldots, \xi_n) := \begin{cases} a_i \xi_i \xi_{i+1}, & i \in \{1, \ldots, n-1\}, \\ a_n \xi_n \xi_1, & i = n. \end{cases}$$

In this case, we can only have $\nu_{i,i+1}(x,y) \neq 0$ if $y \in [-|a_i|, +|a_i|)$, and thus we obtain the same asymptotic behavior as before for $y \in [-a, +a)$, where $a := \min_i |a_i|$, while it changes with $|y|$ increasing until we reach a rate of $1/(\sqrt{2}n)$ when we arrive at $a' := \max_i |a_i|$ (if there is a single i such that $a_i = a'$). In particular, rescaling the functions φ_i leads to the same rate functions as before just for different values of y.

As a slightly more advanced example, consider

$$X_i := \varphi_i(\xi_1, \ldots, \xi_n) := \sum_{j=1}^{n}{}^{(i)} \xi_i \xi_j \tag{38}$$

for all $i = 1, \ldots, n$. (Again, we assume $N = n$.) It is clear that the X_i will take values in $\{-(n-1) + 2m \colon m = 0, 1, \ldots, n-1\}$.

However, in this example we will not arrive at any useful concentration rates anymore. For instance, take $y = n - 3$ and $x = (1, \ldots, 1)$. Then, for any $i \in \{1, \ldots, n\}$ and any $1 \leq k \neq l \leq n$, we have

$$\mathbb{I}(\varphi_i(x) \leq y) - \mathbb{I}(\varphi_i(\sigma_k x) \leq y) - \mathbb{I}(\varphi_i(\sigma_l x) \leq y) + \mathbb{I}(\varphi_i(\sigma_{kl} x) \leq y) = -1$$

and thus $\nu_{kl}(x,y) = -1/4$. This means that

$$\tau(y) = \sup_{x \in \{\pm 1\}^n} \Big(\sum_{1 \leq k \neq l \leq n} \nu_{kl}(x,y)^2 \Big)^{1/2} \geq \sqrt{n(n-1)}/4,$$

a rate which is of course useless.

The reason is that while in (37), the Hoeffding decomposition of $F_n(y)$ with respect to the ξ_i will stop with the second order terms, this does not hold in (38) anymore. As we have seen in Section 8, our results do not perform well if there are higher order Hoeffding terms which dominate the second order term. This is what causes problems here.

Our next goal is to establish conditions which guarantee rates of concentration that are still of interest, i. e. smaller than $\mathcal{O}(1)$. Note that if $\nu_{kl} \neq 0$, we necessarily have $|\nu_{kl}| \geq 1/(4n)$. Thus, even if we have $N = n$ and all the ν_{kl} are of order $\mathcal{O}(1/n)$, we still only get fluctuations of order $\mathcal{O}(1)$ (cf. $\tau(y)$ as defined in Example 11.1).

So we need to control the number of zeros among the ν_{kl}. One idea is the following simple counting argument. Fix some pair of indexes $1 \leq k \neq l \leq N$, and determine how many of the functions φ_i depend on both ξ_k and ξ_l. This can be formulated as follows: For any $i = 1, \ldots, n$, set

$$\zeta_i(k) := \begin{cases} 1, & \text{if there is some } x \in \{\pm 1\}^N \text{ s. th. } \varphi_i(x) \neq \varphi_i(\sigma_k x), \\ 0, & \text{else.} \end{cases}$$

In other words, $\zeta_i(k)$ is 1 if φ_i depends on ξ_k and 0 if not. Then,

$$w(k,l) := \sum_{i=1}^{n} \zeta_i(k) \zeta_i(l)$$

counts the number of the indexes i which both depend on ξ_k and on ξ_l. We therefore get a very simple estimate for $\tau(y)$ as in Example 11.1 by

$$\tau(y) \leq \frac{1}{4n}\Big(\sum_{1\leq k\neq l\leq N} w(k,l)^2\Big)^{1/2} = \frac{1}{\sqrt{8}n}\Big(\sum_{1\leq k<l\leq N} w(k,l)^2\Big)^{1/2}.$$

This is helpful for identifying typical collections of functions $\varphi_1, \ldots, \varphi_n$ for which we will certainly get useful concentration rates when applying Example 11.1. For instance, assume that each of the functions φ_i depends on at most d indexes such that if we compare the sets of indexes for any two $i \neq j$, the number of indexes they have in common is of order $\mathcal{O}(1)$. The d indexes will cause $\binom{d}{2} \approx d^2$ of the numbers $w(k,l)$ to be non-zero, and due to the second assumption these $w(k,l)$ will be of order $\mathcal{O}(1)$.

Putting all together, we get that we can expect $\tau(y)$ to be of order

$$\mathcal{O}\Big(\frac{1}{n}\sqrt{nd^2}\Big) = \mathcal{O}\Big(\frac{d}{\sqrt{n}}\Big).$$

As a point of interest, we note this implies that the highest terms of the Hoeffding decompositions of the φ_i with respect to the ξ_j should be of order less than $\mathcal{O}(n^{1/2})$.

It would be desirable to apply the results from Section 10 to random matrices similarly to the work of S. G. Bobkov and F. Götze [B-G3]. However, this is a harder task. One problem is that we cannot use the results about multilinear polynomials from Section 9 in this context anymore. Therefore we will need to search for different methods for evaluating the condition $|\nabla|\nabla f|| \leq 1$.

12 Random Graphs

As a final application of our results, we now study subgraph counting in Erdős-Rényi random graphs $G(n,p)$. Problems of this type have been widely discussed in the past two decades, for instance by J. H. Kim and V. H. Vu [K-V1], [K-V2], S. Janson and A. Ruciński [J-R], S. Janson, K. Oleszkiewicz and A. Ruciński [J-O-R], S. Chatterjee [C] and B. DeMarco and J. Kahn [D-K]. Our own results are particularly inspired by Section 5.3 in the work of R. Adamczak and P. Wolff [A-W], and as in the latter paper we will focus on counting cycles of fixed length.

First we repeat the basic definitions. An undirected graph (or simply graph) $G = (V,E)$ is a collection of a finite set of vertices V and a set of edges (i. e. subsets of V which consist of two elements) E. Moreover, given G, a graph $H = (V', E')$ is called a subgraph of G if we have $V' \subset V$ and $E' \subset E$.

Given $p \in (0,1)$, the Erdős-Rényi random graph $G = G(n,p)$ is a graph with n vertices which we will simply call $1, \ldots, n$ and whose edges are selected independently such that for any two vertices $i \neq j$, with probability p there is an edge between them.

The problem of subgraph counting can then be described as follows: given any graph H with vertex set $\{1, \dots, k\}$ for some $k \leq n$, we search for the number of copies of H which can be found in the Erdős-Rényi random graph G. As in [A-W], we call this number $Y_H(n,p)$. Of course, $Y_H(n,p)$ is a random variable, and in [A-W], an explicit representation of $Y_H(n,p)$ is given.

We now check how our second order results behave in this context. For this, we use that the edges form a collection of independent Bernoulli variables taking values in $\{0,1\}$. That is, given any two numbers $i \neq j$, we consider the random variable $X_{\{i,j\}}$ which is 1 with probability p and 0 with probability $1-p$. (Note that we do not distinguish between the cases $i < j$ and $j < i$.) Then, the collection of these random variables is just the set of the (random) edges of $G(n,p)$. Now, our aim is to apply the results from Section 9.

As in [A-W], we consider the situation where $H = K_3$, i. e. we count the number of triangles in $G(n,p)$. It is clear we can express this number as

$$Y_{K_3}(n,p) := Y := \sum_{i<j<k} X_{\{i,j\}} X_{\{j,k\}} X_{\{k,i\}}.$$

We now relate this to the conditions we imposed in Section 9.

First, we take the Hoeffding decomposition of Y. For any numbers $i' \neq j'$, we clearly have

$$D_{\{i',j'\}} Y = \sum_{k=1}^{n}{}^{(i',j')} (X_{\{i',j'\}} - p) X_{\{j',k\}} X_{\{k,i'\}}.$$

Moreover, we only have $D_{\{i',j'\}\{i'',j''\}} Y \neq 0$ if $\{i',j'\} \cap \{i'',j''\} \neq \emptyset$. Therefore, considering, say, $D_{\{i',j'\}\{j',k'\}} Y$, we get

$$D_{\{i',j'\}\{j',k'\}} Y = (X_{\{i',j'\}} - p)(X_{\{j',k'\}} - p) X_{\{k',i'\}}$$

and similarly

$$D_{\{i',j'\}\{j',k'\}\{k',i'\}} Y = (X_{\{i',j'\}} - p)(X_{\{j',k'\}} - p)(X_{\{k',i'\}} - p),$$

while all third order differences of different type are 0.

Remembering the proof of Theorem 3.4, by combining these results we obtain

$$\begin{aligned} Y = &\binom{n}{3} p^3 + \sum_{i<j} (n-2) p^2 (X_{\{i,j\}} - p) + \sum_{i<j<k} p\big((X_{\{i,j\}} - p)(X_{\{i,k\}} - p) \\ &+ (X_{\{i,j\}} - p)(X_{\{j,k\}} - p) + (X_{\{i,k\}} - p)(X_{\{j,k\}} - p)\big) \\ &+ \sum_{i<j<k} (X_{\{i,j\}} - p)(X_{\{j,k\}} - p)(X_{\{k,i\}} - p) \end{aligned}$$

as the Hoeffding decomposition of Y. In the second order term, we have combined all types of differences involving three different fixed indexes i, j, k.

Clearly, the $X_{\{i,j\}}$ all have variance $p(1-p)$. We therefore set

$$\psi_{\{i,j\}}(x) := \psi(x) := (x-p)/\sqrt{p(1-p)}$$

for each pair $i \neq j$, so that we have $\mathbb{E}\psi(X_{\{i,j\}}) = 0$ and $\mathbb{E}\psi(X_{\{i,j\}})^2 = 1$ as required in (29). Moreover, an upper bound on $|\psi|$ on $\{0,1\}$ is clearly given by $M := \max(\sqrt{p/(1-p)}, \sqrt{(1-p)/p})$. Now we rewrite the Hoeffding decomposition of Y as

$$\begin{aligned} Y = &\binom{n}{3}p^3 + \sum_{i<j}(n-2)p^{5/2}(1-p)^{1/2}\psi(X_{\{i,j\}}) \\ &+ \sum_{i<j<k} p^2(1-p)\big(\psi(X_{\{i,j\}})\psi(X_{\{i,k\}}) + \psi(X_{\{i,j\}})\psi(X_{\{j,k\}}) + \psi(X_{\{i,k\}})\psi(X_{\{j,k\}})\big) \\ &+ \sum_{i<j<k} p^{3/2}(1-p)^{3/2}\psi(X_{\{i,j\}})\psi(X_{\{j,k\}})\psi(X_{\{k,i\}}), \end{aligned}$$

which is obviously a multilinear polynomial in the $\psi(X_{\{i,j\}})$.

We can thus apply Example 9.2 if we remove the terms of order 0 and 1, i. e. we consider

$$Y - \binom{n}{3}p^3 - \sum_{i<j}(n-2)p^{3/2}(1-p)^{1/2}\psi(X_{\{i,j\}}) =: F.$$

So it remains to deduce suitable values for B_1 and B_2 as in (31) and (32).

For this, we first consider the dediagonalized Hessian of F with respect to D. Again due to (18), this is the same as the dediagonalized Hessian of Y with respect to D, and here we know from above that it has entries

$$\begin{aligned} D_{\{i,j\}\{j,k\}}Y &= (X_{\{i,j\}} - p)(X_{\{j,k\}} - p)X_{\{i,k\}} \\ &= p(1-p)\psi(X_{\{i,j\}})\psi(X_{\{j,k\}})X_{\{i,k\}} \end{aligned}$$

whenever the two sets of indexes $\{i,j\}$ and $\{i',j'\}$ have precisely one element in common (say, j as we just assumed) and entries 0 else.

Remembering (31) and in particular Definition 9.1, we therefore get

$$p(1-p)\sqrt{(\psi(X_{\{i,j\}})^2+1)/2}\sqrt{(\psi(X_{\{j,k\}})^2+1)/2}X_{\{i,k\}} \leq \frac{1}{2}\max(p, 1-p).$$

Here we have used $\sqrt{(\psi(X_{\{i,j\}})^2+1)/2} \leq \max(\sqrt{1/(2(1-p))}, \sqrt{1/(2p)})$. As there are $n(n-1)(n-2)$ matrix entries which are possibly non-zero, we then obtain that we can set

$$B_1 := \frac{1}{2}\sqrt{n(n-1)(n-2)}\max(p, 1-p).$$

Furthermore, a generalization of (26) (note that we also have third order terms in the present situation) yields

$$\begin{aligned}\nabla_{\{i,j\}}F =& \Big|\sum_{k=1}^{n}{}^{(i,j)}\Big(p^2(1-p)(\psi(X_{\{i,k\}})+\psi(X_{\{j,k\}}))\\ &+p^{3/2}(1-p)^{3/2}\psi(X_{\{i,k\}})\psi(X_{\{j,k\}})\Big)\Big|\sqrt{(\psi(X_{\{i,j\}})^2+1)/2}\end{aligned}$$

for any $i \neq j$. Arguing in a similar way as above and putting everything together, we see that we can take

$$B_2 := \frac{3}{\sqrt{2}}(n-2)\max(p,1-p)^{5/2}.$$

Applying Example 9.2 and using $B_1+B_2 \leq 3B_1$, we arrive at the following result:

Example 12.1. *Define $Y_{K_3}(n,p)$ as above, and let*

$$Y_{K_3}(n,p) = \mathbb{E}Y_{K_3}(n,p) + Y_{K_3,1}(n,p) + Y_{K_3,2}(n,p) + Y_{K_3,3}(n,p)$$

be its Hoeffding decomposition with respect to the random variables $X_{\{i,j\}}$, $i<j$. Set

$$F_{K_3}(n,p) := Y_{K_3}(n,p) - \mathbb{E}Y_{K_3}(n,p) - Y_{K_3,1}(n,p).$$

Then, we have

$$\int \exp\left(\frac{2}{15\sqrt{n(n-1)(n-2)}\max(p,1-p)}|F_{K_3}(n,p)|\right)d\mu \leq 2.$$

As a consequence, it follows that

$$\mathbb{P}(|F_{K_3}(n,p)| \geq t) \leq 2\exp(-2t/(15\sqrt{n(n-1)(n-2)}\max(p,1-p)))$$

for any $t>0$.

Here we have replaced the factor 8 in the denominator of the bound given in Example 9.2 by 5 due to Theorem 1.2 and Remark 1.3. Note that in a way, Example 12.1 yields rates of concentration which are uniform in p.

Now we compare our results to the concentration properties of $Y_{K_3}(n,p) - \mathbb{E}Y_{K_3}(n,p)$ as stated in Adamczak and Wolff [A-W], for instance. For example, Proposition 5.5 in [A-W] yields

$$\begin{aligned}&\mathbb{P}(|Y_{K_3}(n,p) - \mathbb{E}Y_{K_3}(n,p)| \geq t)\\ \leq& 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{L_p^6n^3+L_p^4p^2n^3+L_p^2p^4n^4}, \frac{t}{L_p^3n^{1/2}+L_p^2pn}, \frac{t^{2/3}}{L_p^2}\right)\right)\end{aligned} \tag{39}$$

for any $t>0$, where $L_p = (\log(2/p))^{-1/2}$.

To compare this to Example 12.1, we fix some $p > 0$ and consider values of t which are of order $t = \mathcal{O}(n^k)$ for some $k \geq 0$. Then, (39) can roughly be rewritten as

$$\mathbb{P}(|Y_{K_3}(n,p) - \mathbb{E}Y_{K_3}(n,p)| \geq rn^k) \leq 2\exp\left(-C_{r,p}\min\left(n^{2k-4}, n^{k-1}, n^{2k/3}\right)\right)$$

for $r > 0$, while Example 12.1 yields

$$\mathbb{P}(|F_{K_3}(n,p)| \geq rn^k) \leq 2\exp(-C'_{r,p}n^{k-3/2}).$$

It is thus easily seen that our second order results will yield better concentration rates for $k < 5/2$ and $k > 9/2$, hence small and large values for t, where the large values are not really of interest since the maximal number of triangles in $G(n,p)$ is $\binom{n}{3}$ anyway.

On the other hand, the point of view taken in many articles on subgraph counting in $G(n,p)$ is somewhat different. That is, usually the behavior of p (possibly depending on n) is taken into account as well, and the domain of particular interest is the law of large numbers regime. In other words, in the above setting we take $t = \varepsilon\mathbb{E}Y_H(n,p)$ for some $\varepsilon > 0$ fixed and examine how concentration varies in n and p. In general, upper and lower tails, i. e. $\mathbb{P}(\pm(Y_H(n,p) - \mathbb{E}Y_H(n,p)) \geq \varepsilon\mathbb{E}Y_H(n,p))$, are studied separately, and the lower tails yield sharper concentration results than the upper ones.

For instance, one result due to S. Janson, K. Oleszkiewicz and A. Ruciński [J-O-R] is as follows: given $\varepsilon > 0$ such that $\mathbb{P}(Y_H(n,p) - \mathbb{E}Y_H(n,p) \geq \varepsilon\mathbb{E}Y_H(n,p)) > 0$, we have

$$\begin{aligned}\exp\left(-C(H,\varepsilon)M_H^*(n,p)\log\frac{1}{p}\right) &\leq \mathbb{P}(Y_H(n,p) - \mathbb{E}Y_H(n,p) \geq \varepsilon\mathbb{E}Y_H(n,p)) \\ &\leq \exp\left(-c(H,\varepsilon)M_H^*(n,p)\right),\end{aligned}$$

where $c(H,\varepsilon)$ and $C(H,\varepsilon)$ are constants and $M_H^*(n,p)$ is some function (whose explicit expression we skip). Typically, the constants $c(H,\varepsilon)$ and $C(H,\varepsilon)$ are not given much attention.

In the situation we studied above, i. e. $H = K_3$, Corollary 1.7 in [J-O-R] yields that we have $C^{-1}n^2p^2 \leq M_{K_3}^*(n,p) \leq Cn^2p^2$ for some constant C and $p \geq 1/n$. On the other hand, from Example 12.1 it follows that our results lead to

$$\mathbb{P}(|F_{K_3}(n,p)| \geq \varepsilon\mathbb{E}Y_{K_3}(n,p)) \leq 2\exp(-C\varepsilon n^{3/2}p^3/\max(p,1-p))$$

for some absolute constant $C > 0$. That is, for $p \geq 1/2$ we get a rate of $n^{3/2}p^2$, while as $p \to 0$ we only have a rate of $n^{3/2}p^3$, which is unfortunately weaker than the results by [J-O-R] and [A-W]. (In particular, from Theorem 1.5 and Corollary 1.7 in [J-O-R] we know that we should expect a rate of $\Theta(1)$ if $p \leq n^{-1}$, something we clearly do not recover here.)

This is partly because our conditions as in Theorem 1.1 make use of suprema, which leads to estimates as in Example 12.1 which are in some sense uniform in p.

Moreover, even for p fixed we have seen that the law of large numbers regime, i. e. roughly speaking $t = rn^k$ with $k = 3$, is not the domain where our second order concentration results perform best anyway.

Similar observations also hold for cycles of arbitrary length. Without going too much into the details, denote by Y_{K_m} the number of cycles of length m in $G(n,p)$, and set

$$F_{K_m}(n,p) := Y_{K_m}(n,p) - \mathbb{E}Y_{K_m}(n,p) - Y_{K_m,1}(n,p),$$

where $Y_{K_m,1}$ denotes the first order term of the Hoeffding decomposition of Y_{K_m} with respect to the random variables $X_{\{i,j\}}$, $i \leq j$.

Arguing in a similar way as above, we then get a result of the type

$$\mathbb{P}(|F_{K_m}(n,p)| \geq t) \leq 2\exp(-Ct/(n^{m-3/2}\max(p,1-p)))$$

for any $t > 0$, where C is some constant which only depends on m. We can reformulate this as

$$\mathbb{P}(|F_{K_m}(n,p)| \geq \varepsilon\mathbb{E}Y_{K_m}(n,p)) \leq 2\exp(-C\varepsilon n^{3/2}p^m/\max(p,1-p))$$

again with some universal constant $C = C(m)$.

Comparing this to results like Proposition 5.6 in [A-W] or Theorem 1.5 and Corollary 1.7 in [J-O-R], we get similar results as in the case of triangles. That is, for fixed p there are situations (typically for small t or ε small) in which we obtain better concentration rates than the ones obtained in [A-W] and [J-O-R]. On the other hand, as we have to take suprema we get rates which are in some sense uniform in p. In particular, we therefore cannot recover the optimal rate functions $M^*_{K_m}(n,p)$ which study concentration in n and p. Moreover, with m growing larger, we once again observe that our results yield weaker estimates as the influence of higher order Hoeffding terms increases.

A Appendix

In Theorem 1.1, we have required f to be measurable and bounded on the support of μ. A natural idea would be to weaken the boundedness assumption by, for instance, only requiring some moment conditions on f. On the other hand, the function f must fulfill $|\nabla|\nabla f|| \leq 1$, or, more in general, $|\nabla|\nabla f|| < \infty$ on the support of μ. In the sequel, we will show that in fact, this condition already implies the boundedness of f.

For that, we will prove that f is bounded if and only if $|\nabla f|$ is bounded. (Actually, we only need to consider the support of μ, a fact we will ignore from now on.) One direction is trivial. Therefore, it suffices to prove the following lemma:

Lemma A.1. *Let $X_1, \ldots X_n$ be independent random variables with distributions μ_i on $(\mathbb{R}, \mathbb{B})$, $i = 1, \ldots n$, and let $\mu = \otimes_{i=1}^n \mu_i$ be the product measure. Moreover, let $f : \mathbb{R}^n \to \mathbb{R}$ any function in $L^2(\mu)$. Assume that f is not in $L^\infty(\mu)$. Then, $|\nabla f|$ is not in $L^\infty(\mu)$ either.*

Assuming $f \in L^2(\mu)$ is necessary in order to apply the gradient operator ∇. Below, we will give a sketch of the proof but only in a simplified situation. However, it is then pretty clear how the general case will work.

Proof. We assume that all the μ_i are probability measures on $\mathbb{N}$. The proof then works by induction.

So consider $n = 1$ with $\mu_1 \equiv \mu$, and assume that we have $|\nabla f| \leq M$ for some universal constant $M > 0$. We can furthermore assume $\mu(\{1\}) > 0$. Then, we have

$$M \geq |\nabla f(k)| = \big(\frac{1}{2} \sum_{l=1}^{\infty} (f(k) - f(l))^2 \mu(\{l\})\big)^{1/2} \geq |f(k) - f(1)| \sqrt{\mu(\{1\})/2}$$

for all $k \in \mathbb{N}$. However, it follows that f must be bounded, which is a contradiction since we assumed that $f \notin L^\infty(\mu)$.

In the induction step, we consider some function f on $\mathbb{N}^{n+1}$ and once again assume $|\nabla f| \leq M$ for some universal constant M. As above, we can also assume $\mu_{n+1}(\{1\}) > 0$. Now first consider the function

$$f_{[n]} : \mathbb{N}^n \to \mathbb{N}; \qquad (k_1, \ldots, k_n) \mapsto f(k_1, \ldots, k_n, 1).$$

As $|\nabla f| \leq M$, it follows that in particular we have $|\nabla f_{[n]}| \leq M$, and by induction we therefore get that $f_{[n]}$ must be bounded by some universal constant, say, M_1.

To continue, take any vector $(m_1, \ldots, m_n, m_{n+1}) \in \mathbb{N}^{n+1}$ and consider the function

$$f_{n+1} : \mathbb{N} \to \mathbb{R}; \qquad k \mapsto f(m_1, \ldots, m_n, k).$$

Again, in particular we have $|\nabla f_{n+1}| \leq M$. However, as in the case of $n = 1$ we see that we have

$$M \geq |f_{n+1}(k) - f_{n+1}(1)| \sqrt{\mu_{n+1}(\{1\})/2}$$

for all $k \in \mathbb{N}$, and thus, f_{n+1} is bounded by some universal constant M_2 which does not depend of the choice of $m_1, \ldots, m_n$. Combining both arguments leads to

$$\begin{aligned} |f(m_1, \ldots, m_n, m_{n+1})| &\leq |f_{n+1}(m_{n+1}) - f_{n+1}(1)| + |f_{[n]}(m_1, \ldots, m_n)| \\ &\leq M_1 + M_2, \end{aligned}$$

so that f would be bounded by the universal constant $M_1 + M_2$. This once again leads to a contradiction.

It is possible to generalize this proof by choosing suitable partitions of the underlying spaces, for instance. We will then have to work with estimates rather than exact values of f. □

References

[A-W] Adamczak, R., Wolff, P.: *Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order.* Probability Theory Related Fields 162(3) (2015), 531–586.

[A-M-S] Aida, S., Masuda, T., Shikegawa, I.: *Logarithmic Sobolev inequalities and exponential integrability.* J. Func. Anal. 126 (1994), 83–101.

[A] Arcones, M. A.: *A Bernstein-type inequality for U-statistics and U-processes.* Stat. Probab. Letters 22 (1995), 223–230.

[A-G] Arcones, M. A., Giné, E.: *Limit theorems for U-processes.* Ann. Prob. 21 (1993), 1494–1542.

[B-C-G] Bobkov, S. G., Chistyakov, G. P., Götze, F.: *Second Order Concentration on the Sphere.* Preprint, arXiv:1502.04178.

[B-G1] Bobkov, S. G., Götze, F.: *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities.* J. Func. Anal. 163(1) (1999), 1–28.

[B-G2] Bobkov, S. G., Götze, F.: *Concentration inequalities and limit theorems for randomized sums.* Probability Theory Related Fields 137 (2007), 49–81.

[B-G3] Bobkov, S. G., Götze, F.: *Concentration of empirical distribution functions with applications to non-i.i.d. models.* Bernoulli 16(4) (2010), 1385–1414.

[B-G-H] Bobkov, S. G., Götze, F., Houdré, C.: *On Gaussian and Bernoulli covariance representations.* Bernoulli 7(3) (2001), 439–451.

[C] Chatterjee, S.: *The missing log in large deviations for triangle counts.* Random Structures Algorithms 40(4) (2012), 437–451.

[D-G] de la Peña, V., Giné, E.: *Decoupling. From Dependence to Independence.* Springer, 1999.

[D-K] DeMarco, B., Kahn, J.: *Upper tails for triangles.* Random Structures Algorithms 40(4) (2012), 452–459.

[G] Gross, L.: *Logarithmic Sobolev inequalities.* Amer. J. Math 97(4) (1975), 1061–1083.

[H] Hoeffding, W.: *A class of statistics with asymptotically normal distribution.* Ann. Math. Statist. 19 (1948), 293–325.

[J-O-R] Janson, S., Oleszkiewicz, K., Ruciński, A., *Upper tails for subgraph counts in random graphs.* Israel J. Math. 142 (2004), 61–92.

[J-R] Janson, S., Ruciński, A., *The infamous upper tail.* Random Structures Algorithms 20(3) (2002), 317–342.

[K-V1] Kim, J. H., Vu, V. H.: *Concentration of multivariate polynomials and its applications.* Combinatorica 20(3) (2000), 417–434.

[K-V2] Kim, J. H., Vu, V. H.: *Divide and conquer martingales and the number of triangles in a random graph.* Random Structures Algorithms 24(2) (2004), 166–174.

[L1] Ledoux, M.: *On Talagrand's deviation inequalities for product measures.* ESAIM Prob. & Stat. 1 (1996), 63–87.

[L2] Ledoux, M.: *Concentration of measure and logarithmic Sobolev inequalities. Séminaire de Probabilités XXXIII.* Lecture Notes in Math. 1709, 120–216. Springer, 1999.

[L3] Ledoux, M.: *The Concentration of Measure Phenomenon.* American Mathematical Society, 2001.

[M] Major, P.: *On the estimation of multiple random integrals and U-statistics.* Lecture Notes in Math. 2079. Springer, 2013.

[M-O-O] Mossel, E., O'Donnell, R., Oleszkiewicz, K.: *Noise stability of functions with low influences: Invariance and optimality.* Ann. Math. 171 (2010), 295–341.

[R-V] Rudelson, M., Vershynin, R.: *Hanson-Wright inequality and sub-gaussian concentration.* Electronic Communications in Probability 18 (2013), 1–9.

[S-S1] Schudy, W., Sviridenko, M.: *Concentration and Moment Inequalities for Polynomials of Independent Random Variables.* Preprint, arXiv:1104.4997.

[S-S2] Schudy, W., Sviridenko, M.: *Bernstein-like Concentration and Moment Inequalities for Polynomials of Independent Random Variables: Multilinear Case.* Preprint, arXiv:1109.5193.

[T1] Talagrand, M.: *Concentration of measure and isoperimetric inequalities in product spaces.* Publications Mathématiques de l'I.H.E.S. 81(9) (1995), 73–205.

[T2] Talagrand, M.: *New concentration inequalities in product spaces.* Invent. Math. 126 (1996), 505–563.