A Large Deviation Principle for m–variate von Mises–Statistics and U–Statistics

Peter Eichelsbacher , Matthias Löwe
Fakultät für Mathematik, Universität Bielefeld
Postfach 100131, 33501 Bielefeld, Germany
e–mail: loewe@mathematik.uni-bielefeld.de

Abstract

A large deviation principle for m-variate von Mises-statistics and U-statistics with a kernel function satisfying natural moment conditions is proved. Sanov's large deviation result for the empirical distribution function and two fundamental conservation principles in large deviation theory are the main tools. The rate functions are "drawback"-entropy functionals.

AMS Subject Classification (1992): 60F10, 62H10
running title: Large deviations for von Mises and U–Statistics

1 Introduction

Let V be a topological space and $\mathcal{B}(V)$ the Borel σ–field in V. A sequence of probability measures $(Q_n)_{n\geq 1}$ satisfies a large deviation principle (LDP) with rate function $I(\cdot)$ if $Q_n(A)$ can be approximated by $\exp(-n \inf_{x\in A} I(x))$ for appropriate events A in $\mathcal{B}(V)$. We are interested in those events, which describe a deviant behavior of a given ergodic phenomenon.
The following definition due to Varadhan [24] makes this more precise:

A sequence of probability measures $(Q_n)_{n\geq 1}$ on $(V, \mathcal{B}(V))$ obeys a large deviation principle (LDP) with speed a_n and rate function I if there exists a sequence $(a_n)_{n\geq 1}$ of positive numbers tending to infinity and a function $I : V \longrightarrow [0, \infty]$ such that the following conditions hold:

1. $I(x)$ is a lower semicontinuous function. $I \not\equiv 0$.
2. $\{x \ : \ I(x) \leq C\} \subseteq V$ is compact for all $C < +\infty$ (compact level sets).
3. For every open set $O \subseteq V$

$$\liminf_{n\to\infty} \frac{1}{a_n} \log Q_n(O) \geq - \inf_{x\in O} I(x). \tag{1}$$

4. For every closed set $A \subseteq V$

$$\limsup_{n\to\infty} \frac{1}{a_n} \log Q_n(A) \le - \inf_{x\in A} I(x). \tag{2}$$

A sequence $(X_n)_{n\ge 1}$ of random variables with values in V is said to obey a LDP with rate function I satisfying 1. and 2. if the corresponding distributions satisfy 3. and 4.. In this case we will shortly write

$$P(X_n \sim a) \approx e^{-nI(a)}.$$

Already in 1938 Cramér [9] developed the foundations of this theory. He observed that a probability measure induced by the average of n i.i.d. random variables $(X_i)_i$ obeys a principle of large deviations (also see Chernoff [7], Varadhan [24]). 1957 Sanov [23] extended these results to the distribution of the empirical measure of n i.i.d. random variables

$$L_n(\omega, B) = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}(B), \tag{3}$$

where δ denotes the Dirac–measure, $B \in \mathcal{B}(V)$ and $L_n(\omega,\cdot)$ is a random element in $\mathcal{M}^1(V)$, the space of all probability measures on V endowed with the weak topology. In this case the rate function is the relative entropy $H(\varrho|\pi)$ of a probability measure ϱ on V with respect to the distribution π of X_1, where

$$H(\varrho|\pi) := \begin{cases} \int_X f \ \log \ f d\pi & \text{if } \varrho \ll \pi \quad \text{and} \quad f = \frac{d\varrho}{d\pi} \\ \infty & \text{otherwise.} \end{cases} \tag{4}$$

This, roughly speaking, means that under the conditions above the empirical measure converges weakly to the underlying measure exponentially fast.
The key tools in the proofs of Cramér's and Sanov's theorem are basic for the further results, especially the use of Chebyshev inequalities and an exponential change of measure. The large deviation principle found many applications in statistics, in stochastic differential equations, statistical mechanics and the importance sampling method, a special Monte Carlo simulation method.

Donsker and Varadhan intensively studied large deviations for Markov processes [11] as well as for Gaussian processes [12]. As a kind of generalization of the Markov case large deviations for Gibbs measures have been investigated by Föllmer and Orey [15] and by Comets [8]. Also there have been attempts to give large deviation principles in the finest possible topology as in [16] or [17] or to derive more precise estimates as in [5] and [6].
Often a key tool for proving a LDP is to get back to an independent background. Now some of the results below can be summarized by saying that these ideas also work, if we consider a process which is created by independent random variables but itself has got a long range interaction.

To be more precise:
Throughout this paper $(X, \mathcal{B}(X), \pi)$ denotes a probability space and, except otherwise specified, $(\Omega, \mathcal{F}, P)$ is its $\mathbb{N}$-fold product. Define $(X_n)_{n\geq 1}$ to be a sequence of i.i.d. random variables with values in X and distribution (law) π, i.e. for every i $P(X_i \in A) = \pi(A)$.
Let $h : X^m \to \mathbb{R}$ be a measurable and (without loss of generality) symmetric function. The function

$$U_n := U_n(h, X_1, \ldots, X_n) := \frac{1}{\binom{n}{m}} \sum_{1\leq i_1<\ldots<i_m\leq n} h(X_{i_1}, \ldots, X_{i_m}), \quad m \leq n \tag{5}$$

is called U-statistic of degree m with kernel h.

$$V_n := V_n(h, X_1, \ldots, X_n) := \frac{1}{n^m} \sum_{1\leq i_1,\ldots,i_m\leq n} h(X_{i_1}, \ldots, X_{i_m}) \tag{6}$$

is known as the V-statistic or von Mises-statistic (of degree m with kernel h).

After their development by von Mises [22], Halmos [18], and Hoeffding [19], U-statistics as well as von Mises-statistics appeared naturally as a kind of generalized mean value in the theory of statistics (where U_n can be considered as the unbiased variant of V_n). The special attractiveness of U-statistics originates in the following result (cf. [18]): For any unbiased estimator S of a parameter Θ there exists a U-statistic U_n estimating Θ based on the same n observations such that

$$Var\, U \leq Var\, S \tag{7}$$

with equality if and only if $U = S$ a.e. That means U-statistics are best in variance under all unbiased estimators.

Hence it is natural to ask for a large deviation principle for the distribution of such statistics. To this end we use one of the guiding principles in large deviation theory: the invariant nature of large deviation results under mappings which are sufficiently continuous, called the contraction principle (see [10], [24]). In the main theorem of the current paper, this contraction principle is applied to Sanov's theorem.

The organization of this paper is as follows: The main result of this article is formulated in Section 2. Under natural moment conditions on an arbitrary, measurable kernel h we can prove a LDP for the according m–variate U–statistics as well as for m–variate V–statistics. Moreover we give an extension to generalized U–statistics and V–statistics.
Section 3 contains the proofs. First we observe that under the moment conditions U–statistics and V–statistics have identical large deviation behavior in the sense that the LDP holds for the U–statistics U_n with rate function I if and only if it holds for the V–statistics V_n with the same rate. This is an application of the second conservation principle, the so called comparison principle [2], which compares the large deviation behavior of two sequences of probability measures with sufficiently small (Prohorov)-distance.
Next we apply the contraction principle to Sanov's theorem. The contraction map is already

continuous if the kernel is bounded and continuous. We are able to drop continuity and boundedness mainly with the help of Lusin's theorem and a factorization of a U–statistic as a mean of averages of i.i.d variables.
Another charge proving a LDP is using the Cramér–type principle for sequences of i.i.d. random vectors via an appropriate representation of the kernel. A detailed treatment is given in [13] and [21]. We will sketch the main idea of this approach.

Independent of our work, M.A. Arcones [1] has obtained results similar to our theorem in the bivariate case. The method of his proof is different from ours.

2 Main results

In this section we state the main theorem of this paper, a large deviation principle for U-statistics and von Mises-statistics with existing moment generating function.
Before stating the theorem let us agree upon some notation:
Let $\mathbb{P}_\diagdown$ denote the set of all decompositions of the set $\{1, \dots, m\}$ into r subsets $P^1, \dots, P^r$ ($r \in \{1, \dots, m\}$). Set $\mathcal{H}_r$ to be the set of all versions of h depending exactly on r variables only, i.e. $h_r \in \mathcal{H}_r$ iff

$$h_r(y_1, \dots, y_r) = h(x_1, \dots, x_m), \tag{8}$$

such that $x_j = y_i$ if and only if $j \in P^i$ for a partition in $\mathbb{P}_\diagdown$ and $1 \le j \le m, 1 \le i \le r$. We always interpret a function in $\mathcal{H}_r$ as function on X^r rather than a function on X^m. With these notations we are able to formulate the following

Theorem 2.1 *Let (X, π) be a Polish probability space and $h : X^m \to \mathbb{R}$ be a measurable and symmetric function. Assume that for each $r \in \{1, \dots, m\}$ and for every $g \in \mathcal{H}_r$*

$$\int_{X^r} \exp(\Theta ||g(x_1, \dots, x_r)||) d\pi^r(x_1, \dots, x_r) < \infty \qquad \text{for all } \Theta > 0, \tag{9}$$

where π^r denotes the r-fold product measure with marginal distributions π and $||\cdot||$ denotes the usual Euclidean norm on $\mathbb{R}$.
Then, if V_n resp. U_n denotes the associated V-statistics resp. U-statistics with kernel h, for all $a \in \mathbb{R}$ we have

$$P(V_n \sim a) \approx e^{-nI(a)} \tag{10}$$

as well as

$$P(U_n \sim a) \approx e^{-nI(a)}. \tag{11}$$

Here

$$I(a) := \inf_{\substack{\varrho \in \mathcal{M}^1(X) \\ F(\varrho) = a}} H(\varrho | \pi), \tag{12}$$

where $H(\varrho|\pi)$ is the entropy function defined in (4) and $F : \mathcal{M}^1(X) \to \mathbb{R}$ is given by

$$F(\varrho) := \int_{X^m} h(x_1, \ldots, x_m) d\varrho^m(x_1, \ldots, x_m). \tag{13}$$

Remarks 2.2 *1. Note that U_n in the case $m = 1$ is a sum of independent random variables and therefore we obtain Cramér's theorem. The equality between the usual formulation of the rate function in Cramér's theorem as the Legendre transform of the cumulant generating function and our formulation as a drawback-entropy is shown in [14] or [11] (in the case of Banach-space valued variables) or is simply seen from the fact that rate functions are unique.*

2. As a key tool in the proof is the contraction principle the drawback rate $I_1(\cdot)$ is non–negative, lower–semicontinuous and has compact level sets, i.e. $\{y : I_1(y) \leq L\}$ is compact for every $L \geq 0$. Moreover $I_1(y) = 0 \iff y = y_0 := F(\pi)$, because of $H(\varrho|\pi) = 0 \iff \varrho = \pi$ and an argument, using semi–continuity and compactness of the level set of $H(\cdot|\pi)$.

3. Using the fact that the rate function equals zero only at the expected value together with a well known argument given by Ellis [14] making use of the Borel–Cantelli Lemma, we can prove a strong law of large numbers, i.e. $U_n \longrightarrow E_P(U_n)$ P–a.s.. While we do not use the inverse martingale structure of U–statistics, we require rather strong assumptions on the moments of U_n instead.

Remark 2.3 *The notion of U–statistics and V–statistics resp. may be generalized to the case of $k(\geq 2)$ independent samples. For example let $(X_i)_i, (Y_i)_i$ be independent i.i.d. samples with values in Z_1 resp. Z_2 distributed according to μ resp. ν. Let $h : Z_1^{m_1} \times Z_2^{m_2} \to \mathbb{R}$ be a measurable function which is symmetric under permutations of the m_j arguments, $j = 1, 2$. Then for sample sizes (n_1, n_2) with $n_j \geq m_j$, $j = 1, 2$ the generalized U–statistic of degree (m_1, m_2) with kernel h is defined by*

$$U_{n_1,n_2} := \frac{1}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \sum_c h(X_{i_1}, \ldots, X_{i_{m_1}}, Y_{j_1}, \ldots, Y_{j_{m_2}})$$

where the summation $\sum_c$ extends over all $1 \leq i_1 < \cdots < i_{m_1} \leq n_1$ and $1 \leq j_1 < \cdots < j_{m_2} \leq n_2$. If we essentially assume that

$$\lim_{n\to\infty} \frac{n_1}{n} = \lambda,$$

with $n := n_1 + n_2$ and $0 < \lambda < 1$ and an exponential boundedness condition similar to (9), $P \circ U_{n_1,n_2}^{-1}$ satisfies a LDP with speed n and rate function

$$I(a) = \inf_{\substack{\varrho: \\ F(\varrho)=a}} \left[\lambda H(\varrho_1|\mu) + (1-\lambda) H(\varrho_2|\nu)\right],$$

where $\varrho = (\varrho_1, \varrho_2) \in \mathcal{M}^1(Z_1) \times \mathcal{M}^1(Z_2)$ and

$$F(\varrho) = \int_{Z_1^{m_1} \times Z_2^{m_2}} h \quad d(\mu^{m_1} \otimes \nu^{m_2}).$$

Example 2.4 *Let $X = \mathbb{R}$ endowed with Borel-σ-field and let π be any probability measure on $\mathbb{R}$ (e.g. a normalized Lebesgue-measure). Let the kernel be of the form $h(x,y) = \max(x,y)$ or $h(x,y) = |x-y|$ (Gini's mean differences). Then, since these kernels are continuous the associated U-statistics and von Mises-statistics obey a principle of large deviations with rate function*

$$I(y) := \inf_{\substack{\varrho \in \mathcal{M}^1(\mathbb{R}) \\ \mathbb{F}(\varrho) = \curvearrowright}} H(\varrho|\pi),$$

where

$$F(\varrho) := \int_{\mathbb{R}\times\mathbb{R}} h(x,y) d(\varrho \otimes \varrho)(x,y).$$

3 Proofs and a second approach

In this section we give a proof of Theorem 2.1 and sketch a second approach towards a LDP for U–statistics.
As an upper bound for the probability of large deviations usually contains the moment generating function we first bound the moment generating function of U–statistics by using a decomposition due to Hoeffding (see [20], Section 5).

Lemma 3.1 *Let U_n be a $\mathbb{R}^d$–valued U–statistic and denote by $||\cdot||$ the Euclidean norm on $\mathbb{R}^d$. Let $M_t(h) := E(\exp(t||h(x_1,\ldots,x_m)||))$, where E denotes the expected value with respect to the probability P, then*

$$M_t(U_n) := E(\exp(t||U_n||)) \leq (M_{\frac{t}{k}}(h))^k$$

where $k = \left[\frac{n}{m}\right]$.

Proof: Define

$$B(x_1,\ldots,x_n) := \frac{1}{k}\left[h(x_1,\ldots,x_m) + h(x_{m+1},\ldots,x_{2m}) + \cdots + h(x_{km-m+1},\ldots,x_{km})\right].$$

Then according to Hoeffding [20]

$$U_n = \frac{1}{n!}\sum_{\text{per}} B(X_{i_1},\ldots,X_{i_n}),$$

where $\sum\limits_{\text{per}}$ denotes the summation over all $n!$ permutations of $\{i_1,\ldots,i_n\}$. Hence

$$\exp(t||U_n||) = \exp(t||(n!)^{-1}\sum_{\text{per}} B(X_{i_1},\ldots,X_{i_n})||) \leq (n!)^{-1}\sum_{\text{per}} \exp(t||B(X_{i_1},\ldots,X_{i_n})||)$$

by Jensen's inequality. Since $B(\cdot)$ is an average of i.i.d. random variables, we get

$$M_t(U_n) \leq E(\exp(t||B(X_{i_1},\ldots,X_{i_n})||)) = E(\prod_{i=1}^{k} \exp(\frac{t}{k}||h(X_1,\ldots,X_m)||)) = (M_{\frac{t}{k}}(h))^k.$$

$\square$

Proof of Theorem 2.1: The first step of the proof is to show that under condition (9) the distributions of U_n and V_n have identical large deviation behavior. In view of Proposition 2 in [2] we have to show that for all $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log \inf\{\delta > 0, P\circ U_n^{-1}(F) \leq P\circ V_n^{-1}(F^\varepsilon) + \delta \text{ for all closed sets } F \subset \mathbb{R}\} = -\infty.$$

In other words it suffices to prove that

$$\lim_{n\to\infty} \frac{1}{n} \log P(||V_n - U_n|| > \varepsilon) = -\infty,$$

for every positive ε.

Now recognize that V_n consists of a U_n-part and a sum of "diagonal" terms. In fact, using the symmetry of h we can decompose V_n such that

$$V_n = \frac{\binom{n}{m} m!}{n^m} U_n + \frac{1}{n^m} \sum_r h(X_{i_1}, \ldots, X_{i_m}),$$

where the sum $\sum_r$ consists of all realizations of h with at least two of $X_{i_1}, \ldots, X_{i_m}$ being identical, i.e. $h \in \cup_{r=1}^{m-1} \mathcal{H}_r$. Therefore

$$\begin{aligned} P(||V_n - U_n|| > \varepsilon) &\leq P(|1 - \frac{n(n-1)\cdots(n-m+1)}{n^m}|\,||U_n|| > \frac{\varepsilon}{2}) \\ &+ P(\frac{1}{n^m}||\sum_r h(X_{i_1}, \ldots, X_{i_m})|| > \frac{\varepsilon}{2}). \end{aligned}$$

Since $n^m - (n(n-1)\cdots(n-m+1)) \leq Cn^{m-1}$ with a constant C depending on m only we can rewrite the right hand side as

$$P(||V_n - U_n|| > \varepsilon) \leq P(||U_n|| > \frac{n\varepsilon}{2C}) + P(\frac{1}{Cn^{m-1}}||\sum_r h(X_{i_1}, \ldots, X_{i_m})|| > \frac{n\varepsilon}{2C}). \tag{14}$$

Because U_n is an average the first term on the right hand side of (14) can be dominated by

$$n^m P(||h|| > \frac{n\varepsilon}{2C}) \leq n^m \exp(-\Theta n \frac{\varepsilon}{2C}) E_{\pi^m}(\exp(\Theta ||h||)),$$

for any positive Θ by the Chebyshev-Markov-inequality. Hence (because of (9))

$$\lim_{n\to\infty} \frac{1}{n} \log P(||U_n|| > \frac{n\varepsilon}{2C}) = -\infty.$$

The second term on the right in (14) also can be regarded as an average, since $\sum_r$ consists of $n^m - (n(n-1)\cdots(n-m+1))$ summands. On the same reasons as above we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log P(\frac{1}{Cn^{m-1}}||\sum_r h(X_{i_1}, \ldots, X_{i_m})|| > \frac{n\varepsilon}{2C}) = -\infty$$

and thus

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \log P(||V_n - U_n|| > \varepsilon) \leq \\ &\max\Bigg[\lim_{n\to\infty} \frac{1}{n} \log P(||U_n|| > \frac{n\varepsilon}{2C}), \\ &\lim_{n\to\infty} \frac{1}{n} \log P(\frac{1}{Cn^{m-1}}||\sum_r h(X_{i_1}, \ldots, X_{i_m})|| > \frac{n\varepsilon}{2C})\Bigg] = -\infty. \end{aligned} \tag{15}$$

Now we can proceed by proving the desired result for the V-statistics. For this purpose we will apply a general form of the contraction principle given in [10], Lemma 2.1.4 to Sanov's theorem (cf. [10], Theorem 3.2.17).

First of all notice that if we define F as in (13) and L_n as in (3) then

$$F(L_n) = V_n.$$

Unfortunately F in general is not continuous. So we have to approximate F by continuous functions. To this end define for every $L \in (0,\infty)$

$$h_L(x_1,\dots,x_m) := \begin{cases} h(x_1,\dots,x_m) & \text{if } ||h(x_1,\dots,x_m)|| \le L \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

According to Lusin's theorem for every $L \in (0,\infty)$ there exists a bounded, continuous function h_L^c such that $||h_L^c|| \le L$ and

$$\pi^m(||h_L - h_L^c|| \ge \frac{1}{L}) \le e^{-L^2}.$$

Finally define the mappings $F_L : \mathcal{M}^1(X) \to \mathbb{R}$ by

$$F_L(\varrho) := \int_{X^m} h_L^c(x_1,\dots,x_m) d\varrho^m(x_1,\dots,x_m). \tag{17}$$

Note that each F_L is continuous in the weak topology on $\mathcal{M}^1(X)$ since $\varrho \mapsto \varrho^r$ is (see [4], Theorem 3.2). Hence all we have to prove is that F_L converges super-exponentially fast to F in the sense that the following two conditions are satisfied:

$$\lim_{L\to\infty} \sup\{||F_L(\varrho) - F(\varrho)|| : \varrho \in \mathcal{M}^1(X) \text{ with } H(\varrho|\pi) \le C\} = 0 \tag{18}$$

for each $C \in (0,\infty)$ and that

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \left[P \circ L_n^{-1}(\varrho \in \mathcal{M}^1(X) : d(F_L(\varrho), F(\varrho)) \ge \delta)\right] = -\infty \tag{19}$$

for each $\delta \in (0,\infty)$. To check (18) take any level set N of $H(\cdot|\pi)$. We have to show that

$$\lim_{L\to\infty} \sup_{\varrho\in N} || \int (h - h_L^c) d\varrho^m|| = 0.$$

By the triangle inequality we get

$$|| \int (h - h_L^c) d\varrho^m|| \le \int ||h - h_L|| d\varrho^m + \int ||h_L - h_L^c|| d\varrho^m. \tag{20}$$

Since $\varrho \in N$ implies $\varrho^m \in N'$ for a level set N' of $H(\cdot|\pi^m)$ and applying Lemma 5.1 of [11] to $(h - h_L)(\varrho^m)$ yields

$$\lim_{L\to\infty} \sup_{\varrho\in N} \int ||h - h_L|| d\varrho^m = 0.$$

For the second term on the right in (20) we decompose $h_L - h_L^c$ as follows

$$||h_L - h_L^c|| = f_L^{(1)} + f_L^{(2)} \tag{21}$$

where $f_L^{(1)} \leq \frac{1}{L}$ and $\pi^m(f_L^{(2)} \neq 0) \leq e^{-L^2}$, $f_1, f_2 \geq 0$. Clearly

$$\lim_{L\to\infty} \sup_{\varrho\in N} \int f_L^{(1)} d\varrho^m = 0.$$

On the other hand, because of

$$xy \leq e^{\sigma x} + \frac{y}{\sigma} \log \frac{y}{\sigma}$$

for all $x, y \geq 0$ and $\sigma > 0$, we obtain for $\varrho \in N$ and $\sigma \geq 1$:

$$\begin{aligned}
\int f_L^{(2)} d\varrho^m &= \int_{f_L^{(2)}\neq 0} f_L^{(2)} \frac{d\varrho^m}{d\pi^m} d\pi^m \\
&\leq \int_{f_L^{(2)}\neq 0} e^{\sigma f_L^{(2)}} d\pi^m + \frac{1}{\sigma} \int_{X^m} |\log \left(\frac{d\varrho^m}{d\pi^m}\right) / \sigma | d\varrho^m \qquad (22)\\
&\leq \int_{f_L^{(2)}\neq 0} e^{\sigma f_L^{(2)}} d\pi^m + \frac{1}{\sigma} \int_{X^m} |\log \frac{d\varrho^m}{d\pi^m} | d\varrho^m,
\end{aligned}$$

where $\frac{d\varrho^m}{d\pi^m}$ denotes the Radon-Nikodym derivative of ϱ^m with respect to π^m. Since

$$\int_{f_L^{(2)}\neq 0} e^{\sigma f_L^{(2)}} d\pi^m \leq e^{L(2\sigma - L)},$$

which is tending to zero with $L \to \infty$ for every fixed σ and since

$$\lim_{\sigma\to\infty} \sup_{\varrho\in N} \frac{1}{\sigma} \int_{X^m} |\log \frac{d\varrho^m}{d\pi^m}| d\varrho^m = 0$$

by Donsker and Varadhan in [11] proof of Lemma 5.1, condition (18) is satisfied.

For the second condition (19) we have to check that for every $\varepsilon > 0$

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log P(||V_n(h - h_L^c)|| > \varepsilon) = -\infty.$$

Or, by the considerations at the beginning of the proof, equivalently

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log P(||U_n(h - h_L^c)|| > \varepsilon) = -\infty, \qquad (23)$$

where the function in the brackets denotes the kernel of the respective statistics. Of course, one would expect to make use of (9). But the application is not straight forward since U_n consists of "long range dependent" variables. In fact we will use Lemma 3.1 to get along with this fact. We make use of the triangle inequality to conclude

$$\begin{aligned}
&P(||U_n(h - h_L^c)|| > \varepsilon) \leq \qquad (24)\\
&P(U_n(||h - h_L||) > \frac{\varepsilon}{2}) + P(U_n(||h_L - h_L^c||) > \frac{\varepsilon}{2}).
\end{aligned}$$

Now the first term on the right hand side is estimated by Lemma 3.1. From the exponential boundedness of $h(\cdot)$ (9) it follows that for any non negative $t \in \mathbb{R}$:

$$\begin{aligned}
P(||U_n(h - h_L)|| > \varepsilon) &\leq \exp(-t\varepsilon) E(\exp(t||U_n(h - h_L)||)) \qquad (25)\\
&= \exp(-t\varepsilon) M_t(U_n(||h - h_L||)) \\
&\leq \exp(-t\varepsilon)(M_{\frac{t}{k}}(h - h_L))^k.
\end{aligned}$$

where $k = \left[\frac{n}{m}\right]$. Therefore substituting t by tk yields

$$P(||U_n(h - h_L)|| > \varepsilon) \leq \exp(-tk\varepsilon)(M_t(h - h_L))^k,$$

thus

$$\frac{1}{n} \log P(||U_n(h - h_L)|| > \varepsilon) \leq \frac{-tk\varepsilon}{n} + \frac{k}{n} \log M_t(h - h_L),$$

and we conclude

$$\limsup_{n\to\infty} \frac{1}{n} \log P(||U_n(h - h_L)|| > \varepsilon) \leq \frac{-t\varepsilon}{m} + \frac{1}{m} \log M_t(h - h_L).$$

Since $t > 0$ has been arbitrary and $\varepsilon > 0$ we have

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log P(U_n(||h - h_L||) > \varepsilon) = -\infty.$$

The second term on the right in (24) is treated analogously. Recall our decomposition (21) of $h_L - h_L^c$, i.e.

$$||h_L - h_L^c|| = f_L^{(1)} + f_L^{(2)}.$$

Again by the triangle inequality we get

$$P(U_n(||h_L - h_L^c||) > \frac{\varepsilon}{2}) \leq P(U_n(f_L^{(1)}) > \frac{\varepsilon}{4}) + P(U_n(f_L^{(2)}) > \frac{\varepsilon}{4}). \tag{26}$$

Since we can take L large such that $\frac{1}{L} < \frac{\varepsilon}{4}$ it follows

$$P(U_n(f_L^{(1)}) > \frac{\varepsilon}{4}) = 0.$$

It remains to show that

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log P(U_n(||f_L^{(2)}||) > \varepsilon) = -\infty$$

for every positive ε. Exactly the same arguments as above yield

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P(U_n(||f_L^{(2)}||) > \varepsilon) \\
\leq\ & \frac{1}{m}\left(-\Theta\varepsilon + \log E_{\pi^m}(e^{\Theta||f_L^{(2)}||})\right) \\
\leq\ & \frac{1}{m}\left(-\Theta\varepsilon + \log(1 + \exp(L(2\Theta - L)))\right)
\end{aligned} \tag{27}$$

for every positive Θ and ε. Letting first $L \to \infty$ and then $\Theta \to \infty$ proves the desired result. □

Remark 3.2 *The proof of Remark 2.3 follows exactly the same lines, except the additional difficulty that we have to investigate the large deviations for the vector of the empirical measures* $(\frac{1}{n_1}\sum_{i=1}^{n_1} \delta_{X_i}, \frac{1}{n_2}\sum_{i=1}^{n_2} \delta_{Y_i})$. *Now Proposition 2.17 in [3] states that the large deviation properties of the marginal measures carry over to the product measure. Thus* $(\frac{1}{n_1}\sum_{i=1}^{n_1} \delta_{X_i}, \frac{1}{n_2}\sum_{i=1}^{n_2} \delta_{Y_i})$ *obeys a LDP with speed* $n = n_1 + n_2$ *and rate function*

$$\lambda H(\cdot|\mu) + (1 - \lambda)H(\cdot|\nu).$$

Remark 3.3 *Groeneboom et. al. [17] have shown Sanov's theorem in the finer τ-topology which makes integration of bounded measurable functions continuous. It would be attractive to make use of this theorem instead of Sanov's theorem for the weak topology and the approximation via Lusin's theorem, here. In fact this seems rather hard, since it is not known (at least to us) whether the mapping $\varrho \to \varrho^m$, $(m \geq 2)$ is continuous in the τ-topology.*

Next we will sketch the main idea of a second approach proving a LDP. For example, let h be a symmetric polynomial, i.e.

$$h(x_1, \ldots, x_m) := \sum_{k_1,\ldots,k_m=0}^{p} a_{k_1,\ldots,k_m} x_1^{k_1} \cdots x_m^{k_m}, \tag{28}$$

where $k_j \in \mathbb{N}$, $0 \leq k_j \leq p$ and $a_{k_1,\ldots,k_m} = a_{\pi(k_1),\ldots,\pi(k_m)}$ for each permutation π on $\{k_1, \ldots, k_m\}$. Without loss of generality $a_{0,\ldots,0} = 0$ and for at least one set $\{j_1, \ldots, j_m\}$ let $p \in \{j_1, \ldots, j_m\}$ and $a_{j_1,\ldots,j_m} \neq 0$. We decompose V_n into a sum of monomials which makes it possible to "contract" Cramér's theorem as follows: Set $Z_j := (X_j^0, \ldots, X_j^p)$ and $S_n^Z := \frac{1}{n} \sum_{j=1}^{n} Z_j$. Then,

$$V_n = \sum_{k_1,\ldots,k_m=0}^{p} a_{k_1,\ldots,k_m} S^Z_{n,k_1+1} \cdots S^Z_{n,k_m+1},$$

where $S^Z_{n,i} := \frac{1}{n} \sum_{j=1}^{n} X_j^{i-1}$. Hence $V_n = F(S_n^Z)$ and $F : \mathbb{R}^{p+1} \to \mathbb{R}$ is defined by

$$F(x) := \sum_{k_1,\ldots,k_m=0}^{p} a_{k_1,\ldots,k_m} x_{k_1+1} \cdots x_{k_m+1}, \tag{29}$$

where x_{k_j} denotes the k_j–coordinate projection. Some more technical arguments yield the following theorem:

Theorem 3.4 *Let h be a symmetric polynomial with degree g. Suppose*

$$\int_{\mathbb{R}} \exp(\sigma|x|)\pi_{h^s}(dx) < \infty \quad \textit{for all } \sigma > 0, \tag{30}$$

where $s := \frac{g}{p}$ and μ_{h^s} is the distribution of h^s induced by π^m. Then $(P \circ U_n^{-1})_n$ satisfies a LDP with $a_n = n$ and rate function

$$I(y) := \inf\{I(x) \ : \ x \in \mathbb{R}^{g+1} \textit{ and } \ y = F(x)\} \ ; \quad y \in \mathbb{R},$$

where I is the Cramér rate

$$I(x) := \sup_{\lambda \in \mathbb{R}} \{\langle \lambda, x \rangle - C(\lambda)\}. \tag{31}$$

where

$$C(\lambda) := \log \left(\int_{\mathbb{R}} \exp \left[\langle \lambda, x \rangle\right] \mu(dx) \right) \ , \ \lambda \in \mathbb{R}.$$

and F is given by (29).

For a detailed analysis see [13].

Of course this method of representing U_n can be generalized to the infinite dimensional case: Let h be a function such that π^m a.s.

$$h(x_1, \dots, x_m) = \sum_{k_1,\dots,k_m=1}^{\infty} \alpha_{k_1,\dots,k_m} \phi_{k_1}(x_1) \cdots \phi_{k_m}(x_m), \tag{32}$$

where ϕ_l is a real–valued functions on X and the coefficients satisfy the same conditions as in the polynomial case. Here we have to take into consideration the vector

$$\Phi(x) := (\phi_1(x), \phi_2(x), \dots). \qquad x \in X$$

If for every n $S_n = \frac{\Phi(X_1)+\cdots+\Phi(X_n)}{n}$ takes values in a Polish subspace Y of $\mathbb{R}^{\mathbb{N}}$ such that the topology on Y makes projections onto the coordinates continuous, we get the following result:

Theorem 3.5 *Let h be given as in (32) such that the sum converges absolutely. Assume the sequence defined by*

$$G_l(x) := \sum_{k_1,\dots,k_m}^{l} \alpha_{k_1,\dots,k_m} x_{k_1} \cdots x_{k_m}$$

(indices denoting the respective coordinates) converges uniformly to $G : Y \longrightarrow \mathbb{R}$,

$$G(x) := \sum_{k_1,\dots,k_m=1}^{\infty} \alpha_{k_1,\dots k_m} x_{k_l} \cdots x_{k_m},$$

uniformly on the compact subsets of Y. If the distribution of the V–statistic with kernel h obeys a LDP and if the distribution of S_n induced by π satisfies a large deviation principle with rate function $J(\cdot)$, the rate function is given by

$$I(y) := \inf_{\substack{x \in Y: \\ G(x)=y}} J(x). \tag{33}$$

Proving a similar result for U–statistics requires an additional condition to ensure that we are able to handle the diagonal terms. For a detailed treatment and different proofs see [21], [13].

Of course these theorems can be naturally extended to the case of generalized U–statistics. The following examples illustrate applications of Theorem 3.4.

Example 3.6 *An unbiased estimator for the variance is*

$$\hat{\sigma}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{\sum_{j=1}^{n} X_j}{n} \right)^2$$

which can be rewritten as

$$\hat{\sigma}_n = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \frac{1}{2} (X_i - X_j)^2. \tag{34}$$

Thus $\hat{\sigma_n}$ is a U-statistic of degree two with kernel

$$h(x,y) = \frac{1}{2}x^2 - xy + \frac{1}{2}y^2. \tag{35}$$

So if

$$C_{\mu_Z}(\lambda) := \log(\int_{\mathbb{R}^2} \exp\langle\lambda, x\rangle \mu_Z(dx)) < \infty \quad \textit{for all } \lambda \in \mathbb{R}^2,$$

where μ_Z denotes the distribution of the vector $Z = (X_1, X_1^2)$, the sequence σ_n satisfies a LDP with rate function

$$I(y) := \inf_{x\in\mathbb{R}^2 : F(x)=y} I_{\mu_Z}(x),$$

where $F(x) = x_2 - x_1^2$ and $I_{\mu_Z}(x) := \sup_{\lambda\in\mathbb{R}^2}\{\langle\lambda, x\rangle - C_{\mu_Z}(\lambda)\}$.

Example 3.7 *Let $(X_i^{(j)})_{i\geq 1}$ $j = 1, \ldots, k(\geq 2)$ be independent i.i.d. sequences, taking values in $\mathbb{R}^d$, distributed according to μ_j. For each sequence denote by $S_{n_j}^{X^{(j)}}$ the mean $\frac{1}{n_j}\sum_{i=1}^{n_j} X_i^{(j)}$ and set*

$$M_n^\alpha := \sum_{j=1}^k \alpha_j S_{n_j}^{X^{(j)}} \qquad ; \quad \alpha_j \in \mathbb{R}^d. \tag{36}$$

If for each j

$$\int \exp\,[< \lambda, x >]\; \mu_j(dx) < \infty \textit{ for all } \lambda \in \mathbb{R}^d,$$

and if $n_j = \lambda_j n + o(n)$ for each j with $n = \sum_{j=1}^k n_j$ and $\lambda_j > 0$, then $(P \circ (M_n^\alpha)^{-1})_n$ satisfies a LDP with $a_n = n$ and rate function

$$I(a) = \inf_{\substack{(x_1,\ldots,x_k): \\ F(x_1,\ldots,x_k)=a}} \sum_{j=1}^k \lambda_j I_{\mu_j}(x_j), \tag{37}$$

where $F : \mathbb{R}^{kd} \longrightarrow \mathbb{R}^d$ is defined by

$$F(x_1, \ldots, x_k) := \sum_{j=1}^k \alpha_j x_j$$

and I_{μ_j} denotes the Cramér rate.

Finally we would like to thank Peter Major for suggesting the use of Lusin's theorem.

References

[1] Arcones, M.A.: Large deviations for U–Statistics; J. Multivariate Anal., Vol.42, No.2, 299–301 (1992)

[2] Baxter, J.R.; Jain, N.C.: A comparison principle for large deviations; Proceedings of the American Mathematical Society Vol.103, No.4, 1235–1240 (1988)

[3] Bezuidenhout, C.: A large deviations principle for small perturbations of random evolution equations; Ann. Probab. 15, No.2, 646–658 (1987)

[4] Billingsley, P.: Convergence of Probability Measures; Wiley & Sons, New York (1968)

[5] Bolthausen, E.: Laplace approximation for sums of independent random vectors; Probab. Theory Relat. Fields 72, 305–318 (1986)

[6] Bolthausen, E.: Laplace approximation for sums of independent random vectors II; Probab. Theory Relat. Fields 76, 167–206 (1987)

[7] Chernoff,H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations; Ann. of Math. Statist. 23, 493-507 (1952)

[8] Comets,F.: Grandes déviations pour des champs de Gibbs sur $\mathbb{Z}$; C.R.Acad.Sci., Paris, Série 1, 303, 511–514 (1986).

[9] Cramér, H.: On a new limit theorem in the theory of probability; Hermann, Paris (1938)

[10] Deuschel, J.-D.; Stroock, D.W.: Large Deviations; Academic Press 137, San Diego (1989)

[11] Donsker,M.D.; Varadhan,S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time (I,III,IV); Commun.Pure Appl.Math.28, 1-47 (1975); 29, 389-461 (1976); 36, 182-212 (1983)

[12] Donsker, M.D.; Varadhan, S.R.S.: Large deviations for stationary Gaussian processes; Commun. Math. Phys. 97, 187–210 (1985)

[13] Eichelsbacher,P.: Varadhans Prinzip großer Abweichungen für spezielle Statistik-Klassen; Dissertation, Universität Bielefeld (1992)

[14] Ellis,R.S.: Entropy, large deviations and statistical mechanics; Grundlehren der Mathematischen Wissenschaften vol.271, Springer (1985).

[15] Föllmer,H.; Orey,S.: Large deviations for the empirical field of a Gibbs measure; Ann.Probab. 16, 961–977 (1988)

[16] Georgii, H.–O.: Large deviations and maximum entropy principle for interacting random fields on $\mathbb{Z}^d$; Preprint München (1991)

[17] Groeneboom, P.; Oosterhoff, J.; Ruymgaart, F.H.: Large deviation theorems for empirical probability measures; Ann. Probab. 7, 553–586 (1979)

[18] Halmos, P.R.: The theory of unbiased estimation; Ann. of Math. Stat. 17, 34–43 (1946)

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

[20] Hoeffding, W.: Probability inequalities for sums of bounded random variables; Journal of the American Statistical Association 58, 13-30 (1963)

[21] Löwe, M.: Exponential Inequalities and Principles of Large Deviations for U–Statistics and von Mises–Statistics; Dissertation; Universität Bielefeld (1992)

[22] von Mises, R.: On the asymptotic distribution of differentiable statistical functions; Ann. of Math. Stat. 18, 309–348 (1947)

[23] Sanov, J.N.: On the probability of large deviations of random variables; Selected Translations in Mathematical Statistics and Probability I, 214–244 (1957)

[24] Varadhan, S.R.S.: Large Deviations and Applications; École d'Été de Probabilités de Saint-Flours XV-XVII, Lecture Notes in Mathematics, Springer (1985–87)