Probab. Theory Relat. Fields 104, 1–14 (1996)

Probability Theory and Related Fields

Refinements of the Gibbs conditioning principle

A. Dembo[1,★], **O. Zeitouni**[2,★★]

[1] Department of Mathematics and Department of Statistics, Stanford University, Stanford, California, USA
[2] Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel

Received: 10 April 1994 / In revised form: 8 May 1995

Summary. Refinements of Sanov's large deviations theorem lead via Csiszár's information theoretic identity to refinements of the Gibbs conditioning principle which are valid for blocks whose length increase with the length of the conditioning sequence. Sharp bounds on the growth of the block length with the length of the conditioning sequence are derived.

Mathematics Subject Classification (1991): 60F10

1 Introduction

Throughout this paper, $X_1, X_2, \ldots$ denotes a sequence of independent, identically distributed random variables, distributed over a Polish space $(\Sigma, \mathscr{B}_\Sigma)$ with common distribution P_X. Here, $\mathscr{B}_\Sigma$ denotes the Borel σ-field of Σ. Let $L_n = \frac{1}{n}\Sigma_{i=1}^n \delta_{X_i}$ denote the empirical measure of the sequence $\{X_i\}_{i=1}^n$, and for any two measures μ, ν, let $H(\mu|\nu)$ denote the relative entropy of μ with respect to ν.

A common situation is the following. One is given an observation of the empirical measure (usually, in the form of some averaged "energy"; for precise definitions, see Sect. 2). One wishes then to deduce information about the distribution of the random sample conditioned on this observation.

The simplest situation in which such a set up occurs is in the "Gibbs conditioning principle" of statistical mechanics. Let $A(a,\delta) = \{\omega : n^{-1}\Sigma_{i=1}^n f(X_i) \in [a-\delta, a+\delta]\}$, for some measurable function $f(\cdot)$. Under suitable conditions on P_X and $f(\cdot)$, the Gibbs conditioning principle is the statement that, for

★ Partially supported by NSF DMS92-09712 grant and by a US–Israel BSF grant
★★ Partially supported by a US–Israel BSF grant and by the fund for promotion of research at the Technion

any Borel set $B \in \mathscr{B}_\Sigma$, as soon as $E_{P_X}(f) \neq a$, one has

$$\lim_{\delta \to 0} \lim_{n \to \infty} P_X^n(X_1 \in B \,|\, A(a, \delta)) = \gamma^*(B) \tag{1.1}$$

where γ^* minimizes the relative entropy $H(\,\cdot\,|P_X)$ under an energy constraint, and satisfies $d\gamma^*(x)/dP_X(x) = e^{\beta f(x)}/Z_\beta$, with $Z_\beta = \int_\Sigma e^{\beta f(x)} P_X(dx)$ and $\beta = \beta(a)$ is chosen such that $E_{\gamma^*}(f) = a$. (For precise statements in this direction, see, e.g., [3, 5, 13, 16]).

Statements of the form (1.1) are a particular case of what we refer to as the "Gibbs conditioning principle", which is the meta-theorem which under the conditioning that the empirical measure belongs to some "rare set" A, the law of X_1 converges to the law which minimizes the relative entropy subject to the constraint of belonging to A. There exist a few approaches to the derivation of such principles. For some remarks on the history of the problem, see the introduction section in [16]. One of the most successful solutions to this question is via the theory of large deviations. Indeed, Gibbs conditioning served as a motivation behind Lanford's subadditive approach to the theory of large deviations. Using the latter, one typically obtains weak convergence of the conditional measure appearing in (1.1) to γ^*, and one may also extend the statement (1.1) to the statement that the law of $X_1, \ldots, X_k$ under the previous energy constraint converges weakly to $(\gamma^*)^k$, with k fixed, namely,

$$\lim_{\delta \to 0} \lim_{n \to \infty} P_X^n((X_1, \ldots, X_k) \in C | A(a, \delta)) = (\gamma^*)^k(C) \tag{1.2}$$

for any $C \in \mathscr{B}_\Sigma^k$. One may also obtain Markov analogues of these results (see, e.g., [3, 6, 15, 16]).

In fact, statements like (1.2) hold for quite general type of constraints, for appropriate γ^* which solve the variational problem of minimizing the relative entropy subject to the constraint. When convexity is present in the constraint (like in the case of energy conditioning described above), a combination of large deviations ideas with geometrical analysis allows Csiszár (see [5]) to obtain a much stronger mode of convergence. Namely, he proves that the convergence is actually in divergence, which implies convergence in variation norm. These results have been extended to the Markov case by Schroeder [15].

Our goal in this paper is to obtain extensions of (1.2) which allow for a growth of k with n, in the case that some convexity is available in the conditioning. In physical terms, this means that one is interested not in the behavior of individual "particles" under the conditioning but rather in the behavior of increasingly large "sub-systems". Obviously, $k(n)$ cannot grow too fast (in particular, one cannot have $k(n) = n$ and still hope to have (1.2). Actually, we show in Proposition 2.12 that, under mild conditions, in order for (1.2) to hold, it is necessary that $k(n) = o(n)$). Our approach to finding growth rates of $k(n)$ which preserve (1.2) is based on the observation that Csiszár's results may be extended to deal with increasing $k = k(n)$ as soon as one has refinements of Sanov's theorem. It seems beyond hope to be able to obtain such refinements in full generality. On the other hand, such refinements are available (with some efforts) in several particular (important) cases, and lead to the corresponding extensions of the Gibbs conditioning principle. The particular examples in Sect. 2 are intended to serve as an illustration to this general phenomenon. A corollary of our results (see Corollary 2.7 for the precise conditions) is that (1.2) remains valid (in the sense of convergence in variation norm) if $k(n)\log n/n \to_{n\to 0} 0$ and, under additional restrictions, as soon as $k(n) = o(n)$, the sharpest rate possible (see Proposition 2.15). Similar results hold for the case with interaction (where the conditioning is with respect to U-statistics, namely the energy is described by a quadratic form involving pairs of points in the sample $X_1,\ldots,X_n$, see Corollary 2.11). These results form the core of Sect. 2. For the sake of better readability, we have postponed many proofs which interrupt the flow of the presentation in Sect. 2 to a separate section.

We remark that Bolthausen [2] has results related to the refinements obtained in this work. However, he works under smoothness assumptions which are not satisfied here, and it is not clear how to extend his results to our setup.

We conclude this introduction with some comments and open problems. First we note that our bounds are not always optimal, and it is of interest to find the maximal rate of growth of $k(n)$ which still yields conditional independence. (Note that even in the simplest situation treated in Corollary 2.7, the gap between the rate of growth of $k(n)$ and the necessary condition of Proposition 2.12 is closed in Proposition 2.15 only under special conditions). This gap is even larger when one considers truly "infinite dimensional" conditioning by a countable number of constraints: there, under appropriate conditions, one has conditional independence with $k(n) = o(\sqrt{n})$ (details available from the authors), but it is not clear whether this is the optimal rate. Next, by an extension of Csiszár's triangle inequality and information theoretic identity, the results extend in a straightforward manner to the case of Markov chains. It should carry over to Markov random fields with local interaction, but we do not carry through this extension. Finally, conditions for the applicability of Proposition 2.8 for general conditioning sets are needed. It is expected that such conditions could be derived based on the yet unavailable local CLT's for empirical measures, thus motivating further study of the latter.

2 Conditioning, and refinements of Sanov's theorem

Let $B(\Sigma), C_b(\Sigma)$ denote the space of bounded measurable (respectively, bounded continuous) functions on Σ. Let $M_1(\Sigma)$ denote the space of probability measures on Σ, equipped with the weak ($C_b(\Sigma)$-) topology which makes it into a Polish space. Recall that a set $\Pi \subset M_1(\Sigma)$ is *completely convex* if for every probability space $(\Omega, \mathscr{B}, \mu)$ and Markov kernel v from $(\Omega, \mathscr{B})$ to $(\Sigma, \mathscr{B}_\Sigma)$ such that $v(\omega, \cdot\,) \in \Pi$ for each $\omega \in \Omega$, the probability measure μv defined by $\mu v(\,\cdot\,) = \int v(x, \cdot\,)\mu(dx)$ also belongs to Π (see [5, Definition 2.3]). A convex set $\Pi \subset M_1(\Sigma)$ is *almost completely convex* if there exists a monotone increasing sequence Π_k of completely convex subsets of Π such that every atomic $v \in \Pi$ with a finite number of atoms is also in $\cup_k \Pi_k$.

For any measure Q, let Q^n denote the n-fold product of Q. We use $P_{L_n} \in M_1(M_1(\Sigma))$ to denote the law of the empirical measure $L_n = \frac{1}{n}\Sigma_{i=1}^n \delta_{X_i}$ in $M_1(\Sigma)$. Whenever $P_{L_n}(\Pi') > 0$, let $P^n_{X^k|\Pi'}$ denote the law of $(X_1,\ldots,X_k)$ conditioned on the event $L_n \in \Pi'$ (here, $k \leq n$).

We shall make throughout the following assumption.

Assumption (A-1) *Π' is a measurable subset of an almost completely convex $\Pi \subset M_1(\Sigma)$, with $P_{L_n}(\Pi') > 0$ and $H(\Pi \mid P_X) = \inf_{P\in\Pi} H(P \mid P_X) < \infty$.*

Here, Π' measurable means that $\{(x_1,\dots,x_n) : n^{-1}\Sigma_{i=1}^n \delta_{x_i} \in \Pi'\} \in \mathscr{B}_{\Sigma^n}(= (\mathscr{B}_\Sigma)^n)$ for all n.

Let P^* be the generalized I-projection of P_X on Π. That is, P^* is the unique element of $M_1(\Sigma)$ such that if $P_m \in \Pi$ satisfy

$$H(P_m|P_X) \to_{m\to\infty} \inf_{P\in\Pi} H(P|P_X)$$

then $P_m \to P^*$, with the convergence holding in variational norm (see [4] for a proof of the existence and uniqueness of the generalized I-projection, and note that if Π is variation-closed then $P^* \in \Pi$ and $H(\Pi|P_X) = H(P^*|P_X)$ by [5, (1.6)]). Henceforth we let $\overline{f} = \log \frac{dP^*}{dP_X} - H(P^*|P_X)$, so that $\overline{f} \in L_1(P^*)$ with $\int \overline{f}\, dP^* = 0$.

A remarkable observation of Csiszár is the following

Theorem 2.1 [5, Theorem 1] *Assume* (A-1). *Then,*

$$\frac{1}{n} H(P^n_{X^n|\Pi'}|(P^*)^n) \leqq -\frac{1}{n} \log P_{L_n}(\Pi') - H(\Pi|P_X)$$

$$\leqq -\frac{1}{n} \log P_{L_n}(\Pi') - H(P^*|P_X)\,. \tag{2.2}$$

In particular, since for any $\mu \in M_1(\Sigma^n)$, $\nu \in M_1(\Sigma)$, with μ_i denoting the marginal of μ on the ith coordinate (see [5, (2.10)]),

$$\frac{1}{n}\sum_{i=1}^{n} H(\mu_i|\nu) \leqq \frac{1}{n} H(\mu|\nu^n)\,. \tag{2.3}$$

Csiszár, using the exchangeability of the random variables X_i under the conditioning, concludes that

$$H(P^n_{X^1|\Pi'}|P^*) \leqq -\frac{1}{n} \log P_{L_n}(\Pi') - H(\Pi|P_X)\,. \tag{2.4}$$

We say that Π' satisfies the Sanov property (with respect to P_X and Π) as soon as the right-hand side of (2.4) converges to zero. In particular, if $\Pi' = \Pi$ is a closed set with nonempty interior such that $\inf_{P\in\Pi^o} H(P|P_X) = \inf_{P\in\Pi} H(P|P_X)$, the Sanov property is a direct consequence of Sanov's theorem. Whenever the Sanov property holds, one obtains the convergence of the conditional measure of X_1 to the measure P^*, in a divergence sense which is even stronger than convergence in variational norm.

Our starting point is the following, well-known refinement of (2.3). Let $n, k(n)$ be such that $n/k(n)$ is an integer. Consider blocks of length $k(n)$, and for any $\mu \in M_1(\Sigma^n)$, denote by $\mu_j^{k(n)} \in M_1(\Sigma^{k(n)})$ the law of the jth block (that is, $\mu_1^{k(n)}(A) = \mu(A \times \Sigma^{n-k(n)})$, and in general $\mu_j^{k(n)}(A) = \mu(\Sigma^{(j-1)k(n)} \times A \times$

$\Sigma^{n-jk(n)})$ for every Borel set $A \subset \Sigma^{k(n)})$. Then, (2.3) reads

$$\sum_{j=1}^{n/k(n)} H(\mu_j^{k(n)}|v^{k(n)}) \leqq H(\mu|v^n). \tag{2.5}$$

Again by the exchangeability of the $k(n)$ blocks, it follows that when (A-1) holds

$$H(P^n_{X^{k(n)}|\Pi'}|(P^*)^{k(n)}) \leqq k(n)\left(-\frac{1}{n}\log P_{L_n}(\Pi') - H(\Pi|P_X)\right)$$
$$\leqq k(n)\left(-\frac{1}{n}\log P_{L_n}(\Pi') - H(P^*|P_X)\right). \tag{2.6}$$

Csiszár has actually observed (2.6) for $k(n) = k$ independent of n, and in this context concluded that as soon as the Sanov property holds, any fixed number of variables X_i behave, under the conditioning, like independent random variables. Note, however, that more information is contained in (2.6): namely, whenever one may prove refinements of Sanov's property, one immediately obtains independent-like behavior for blocks of length related to the accuracy of the refinement. Our goal therefore in this section is to present several situations where such refinements may be obtained, leading to a "Gibbs" statement for n-dependent blocks.

The following simple corollary of (2.6) applies to the conditioning on the empirical mean of $\mathbb{R}^\ell$-valued statistics, i.e., conditioning on the event $\{n^{-1}\Sigma_{i=1}^n \psi(X_i) \in C\}$, where $\psi : \Sigma \to \mathbb{R}^\ell$ is a Borel measurable map. Let $Q_X = P_X \circ \psi^{-1}$, and $\Lambda(\lambda) = \log \int e^{\langle \lambda, x\rangle} Q_X(dx)$.

Corollary 2.7 *Let $\Pi' = \{v : v \circ \psi^{-1}$ of compact support, $\int \psi\, dv \in C\}$, for a convex set $C \subset \mathbb{R}^\ell$ such that C^o intersects the interior of the convex hull of the support of Q_X. Suppose further that Q_X is either lattice or strongly nonlattice, that $\{\lambda : \Lambda(\lambda) < \infty\}$ is an open set, and $\int xQ_X(dx) \notin C$. If $n^{-1}k(n)\log n \to 0$, then $H(P^n_{X^{k(n)}|\Pi'}|(P^*)^{k(n)}) \to 0$.*

(Strongly nonlattice means that the modulus of the Fourier transform of Q_X equals one only at the origin).

Remark. It is shown in [5, (2.36)] that in this setting $\frac{dP^*}{dP_X} = \exp(\langle \lambda^*, \psi(\,\cdot\,)\rangle - \Lambda(\lambda^*))$ where $\lambda^* \in \mathbb{R}^\ell$ attains the maximum of $h(\lambda) = \inf_{x\in C}\langle \lambda, x\rangle - \Lambda(\lambda)$.

Proof. See Sect. 3. □

The rate in Corollary 2.7 is in general not optimal. As will be shown below (cf. Proposition 2.12), $k(n) = o(n)$ is necessary for the conclusion of Corollary 2.7. Under additional assumption, it can be shown (by a somewhat different technique), that $o(n)$ is actually sufficient (see Proposition 2.15).

Since the method of obtaining refinements based on (2.6) is relatively simple to apply, it is of interest to note that in general (2.6) is not tight even when (2.2) is. Consider $P_X = Q_X$, the standard Normal law on $\Sigma = \mathbb{R}$, with $C = [1, \infty)$ and $\psi(\,\cdot\,)$ being the identity map. In this setting P^* is the law of a Normal (1,1) random variable, and the event $L_n \in \Pi'$ corresponds to conditioning on $n^{-1}\Sigma_{i=1}^n X_i \geqq 1$. Using the special structure of the Normal law, it follows as in Proposition 2.15 that $k(n) = o(n)$ suffices for $k(n)$-independence. A direct computation reveals that the difference between the right-side and left-side of (2.2) is at most $1/n$ but for $k(n)/n \to 0$ while $n^{-1}k(n)\log n \to \infty$ the right-side of (2.6) is unbounded yet the left-side of (2.6) converges to zero. The cause of this lies in (2.5) where we ignored the contribution due to the conditional dependence among the $k(n)$-blocks.

In [9, Theorem 1.6], Diaconis and Freedman deal with point conditioning, as in the above example when $C = \{1\}$, and prove that then $H(P^n_{X^k|\Pi'}|(P^*)^k) \to 0$ iff $k(n)/n \to 0$ (their results are phrased in terms of the variation norm, but the estimate of [9, Lemma 3.1] suffices for convergence in divergence). In the setting of [9], $P^n_{X^n|\Pi'} = (P^*)^n_{X^n|\Pi'}$ by sufficiency theory for exponential families, allowing one to let $P_X = P^*$ to begin with. On the other hand, $P^n_{X^n|\Pi'}$ is then singular making (2.2) useless. In contrast, for $C = [1, \infty)$ in the above example $H((P^*)^n_{X^n|\Pi'}|(P^*)^n) \leqq \log 2$ by (2.2), demonstrating the dependence of the conditional distribution on the parameter of the relevant exponential family.

For arbitrary measurable set Π', Diaconis and Freedman show in [8, Theorem 13] that the variational distance between $P^n_{X^n|\Pi'}$ and the set of mixture laws $\{Q \in M_1(\Sigma^k) : Q(\,\cdot\,) = \int P^k(\,\cdot\,)\mu_n(dP),\ \ \mu_n \in M_1(M_1(\Sigma))\}$, is at most k^2/n, and in [8, Proposition 31] give an example of Π' for which this rate is tight. In comparison, our results deal with stronger notion of divergence distance, with $\mu_n = \delta_{P^*}$ which is degenerate and independent of n, but cover only some special classes of sets Π' where typically a much better convergence rate is achievable.

The next proposition is suitable for analyzing the more general setting not covered in Corollary 2.7. As mentioned before, the required tool is a refined lower bound on $P_{L_n}(\Pi')$. The main idea is to perform a change of measure in the proof of the large deviations lower bound to a point which may be an interior point, but which converges with n to a boundary point. This allows to have a ball wholly contained inside the conditioning set, and hence to avoid the need for "local" results, which are generally cumbersome and known only in finite dimensions. On the other hand, this procedure introduces a discrepancy in the exponent which needs to be controlled. Our main application, in this paper, of Proposition 2.8 uses $\alpha_n = 0$.

Proposition 2.8 *Assume* (A-1). *Suppose that for some* $Q \in M_1(\Sigma)$ *with* $H(Q|P_X) < \infty$ *there exist* $\alpha_n \in [0,1]$, $\rho_n > 0$ *and* $k(n)$ *such that* $k(n)(\alpha_n + \rho_n) \to 0$ *and*

$$\liminf_{n\to\infty} \frac{k(n)}{n} \log \left[(Q_{\alpha_n})^n \left(L_n \in \Pi', n^{-1}\sum_{i=1}^n (f_{\alpha_n}(X_i) - \textstyle\int f_{\alpha_n}\, dQ_{\alpha_n}) < \rho_n \right) \right] = 0 \tag{2.9}$$

where $Q_\alpha = \alpha Q + (1-\alpha)P^*$ *and* $f_\alpha = \log \frac{dQ_\alpha}{dP_X}$. *Then,* $H(P^n_{X^{k(n)}|\Pi'}|(P^*)^{k(n)}) \to 0$.

Proof. Fix Q as in the statement of the proposition, and observe that since $H(Q|P_X) < \infty$, by convexity of the relative entropy

$$H(Q_\alpha|P_X) \leqq \alpha H(Q|P_X) + (1-\alpha)H(P^*|P_X) < \infty\,,$$

for all $\alpha \in [0,1]$, so that $f_\alpha = \log \frac{dQ_\alpha}{dP_X} \in L_1(Q_\alpha)$ with $\int f_\alpha \, dQ_\alpha = H(Q_\alpha|P_X)$. Fix any measurable representation of f_α in $L_1(Q_\alpha)$ and let

$$\Pi_{\rho,\alpha} = \{\nu : f_\alpha \in L_1(\nu), \quad \int f_\alpha \, d\nu - \int f_\alpha \, dQ_\alpha < \rho\} \cap \Pi'\,.$$

(Although $\Pi_{\rho,\alpha}$ may depend on the particular representation of f_α chosen in its definition, $(Q_\alpha)_{L_n}(\Pi_{\rho,\alpha})$ does not).

Observe that for every n, $\alpha \in [0,1]$ and $\rho > 0$

$$\begin{aligned}
P_{L_n}(\Pi') \geqq P_{L_n}(\Pi_{\rho,\alpha}) &\geqq \exp[-n(\rho + \textstyle\int f_\alpha \, dQ_\alpha)] \displaystyle\int_{\Pi_{\rho,\alpha}} \exp[n \textstyle\int f_\alpha \, dL_n] \, dP_{L_n} \\
&= \exp[-n(\rho + H(Q_\alpha|P_X))] \int_{\Pi_{\rho,\alpha}} d(Q_\alpha)_{L_n} \\
&\geqq \exp[-nH(P^*|P_X)] \exp[-n(\rho + \alpha H(Q|P_X))] \\
&\quad \times (Q_\alpha)_{L_n}(\Pi_{\rho,\alpha})\,. \qquad (2.10)
\end{aligned}$$

Since $f_\alpha \in L_1(L_n)$ a.e.-$(Q_\alpha)^n$, the proof is complete by combining (2.6), (2.10) and our assumptions on α_n, ρ_n and $k(n)$. □

Remark. In the case $\alpha_n = 0$, (2.10) is related to [10, Theorem 2.1].

Proposition 2.8 applies in the following special case of conditioning by a U-statistics. Let $U : \Sigma^2 \to [0, M]$ be a continuous, symmetric, bounded function, such that:

(C-1) $\int U(x,y)(Q_1 - Q_2)(dx)(Q_1 - Q_2)(dy) \geqq 0$ for every $Q_1, Q_2 \in M_1(\Sigma)$.

(C-2) $\int U(x,y) P_X(dx) P_X(dy) > 1$.

(C-3) There exists $Q \in M_1(\Sigma)$ such that $H(Q|P_X) < \infty$ and $\int U(x,y) Q(dx)Q(dy) < 1$.

Corollary 2.11 *Assume that* (C-1)–(C-3) *hold. Let* $\Pi = \Pi' = \{\nu : \int U(x,y) \nu(dx)\nu(dy) \leqq 1\}$. *Then,* $H(P^n_{X^{k(n)}|\Pi'}|(P^*)^{k(n)}) \to 0$ *provided that* $n^{-1}k(n)\log n \to 0$.

Remark. The conditioning $L_n \in \Pi'$ corresponds to $n^{-2}\Sigma^n_{i,j=1} U(X_i, X_j) \leqq 1$.

Proof. See Sect. 3.

Having spent some effort in obtaining convergence statements for the conditional law $P^n_{X^{k(n)}|\Pi'}$, we next show that under mild conditions, $k(n) = o(n)$ is *necessary* for $H(P^n_{X^{k(n)}|\Pi'}|(P^*)^{k(n)}) \to 0$.

Proposition 2.12 *Let Π be convex and such that $H(\Pi|P_X) < \infty$. Let P^* denote the generalized I-projection of P_X on Π with $\overline{f} = \log(dP^*/dP_X) - H(P^*|P_X)$. Assume that $\overline{f} \in L_2(P^*)$, that*

$$\log P_{L_n}(\Pi') + nH(P^*|P_X) \geqq -\mu(n) \qquad (2.13)$$

for some positive sequence $\mu(n) = o(\sqrt{n})$*, and that the characteristic function of* $P^* \circ \overline{f}^{-1}$ *is in* $L_p(\mathbb{R})$*, some* $p \in [1, \infty)$*. Then, for* $k(n) = \beta n$*, any* $1 > \beta > 0$ *fixed, one has*

$$\liminf_{n\to\infty} H(P^n_{X^{k(n)}|\Pi'}|(P^*)^{k(n)}) > 0\,. \tag{2.14}$$

Remark. (2.13) holds, in particular in the setting of Corollaries 2.7 and 2.11.

Proof. See Sect. 3.

In the situation described by Corollary 2.7, one may under further assumptions actually close the gap between the sufficient rate $k(n) = o(n/\log n)$ and the necessary rate $k(n) = o(n)$. Indeed, we have the following result.

Proposition 2.15 *In the setup of Corollary 2.7, assume that the characteristic function of* $P^* \circ \psi^{-1}$ *is in* $L_p(\mathbb{R}^\ell)$*; some* $p \in [1, \infty)$*. Further assume that, for some* $M < \infty$,

$$P_{L_n}(\Pi')e^{nH(P^*|P_X)} \geqq \frac{n^{-1/2}}{M}\,. \tag{2.16}$$

Then, for any $k(n) = o(n)$,

$$\|P^n_{X^{k(n)}|\Pi'} - (P^*)^{k(n)}\|_{\mathrm{var}} \underset{n\to\infty}{\longrightarrow} 0\,. \tag{2.17}$$

Proof. See Sect. 3.

Remarks. (1) Conditions for (2.16) to hold are given in [11, 14]. In particular, (2.16) holds for $\ell = 1$ and, for $\ell > 1$, as soon as Π' is a convex polytope with P^* belonging to the relative interior of an $(\ell - 1)$-dimensional facet.

(2) If Q_X possesses a bounded density, then the characteristic function of $P^* \circ \psi^{-1}$ is in $L_p(\mathbb{R}^\ell)$ for some $p \in [1, \infty)$.

(3) In Proposition 2.15 we may find other assumptions replacing (2.16) (cf. Remark 3.1).

Proof. See Sect. 3.

3 Proofs

Proof of Corollary 2.7. Let $I(z) = \sup_{\lambda \in \mathbb{R}^\ell} [\langle \lambda, z\rangle - \Lambda(\lambda)]$. Note that $P_{L_n}(\Pi') = Q^n_X(n^{-1}\Sigma^n_{i=1} Y_i \in C)$, where $Y_i = \psi(X_i)$ are i.i.d. Q_X. It follows from [14, (3.4)] that for some finite $c_1 > 0$ and n large enough

$$n^{-1} \log P_{L_n}(\Pi') \geqq \eta + n^{-1} \log(c_1 n^{-\ell/2})\,,$$

where

$$\eta = \lim_{n\to\infty} \frac{1}{n} \log Q^n_X \left(n^{-1} \sum_{i=1}^n Y_i \in C\right) \geqq -\inf_{z \in C^o} I(z) > -\infty\,,$$

and the inequalities follow from Cramèr's theorem and the support condition on Q_X.

In [5, (3.36) and Lemma 4.3], it is shown that $\Pi = \{\nu : \nu \circ \psi^{-1}$ of compact support, $\int \psi \, d\nu \in \overline{C}\}$, is almost completely convex. As $n^{-1}k(n)\log(c_1 n^{-\ell/2}) \to 0$, the proof is completed by (2.6) provided that $H(\Pi/P_X) \geqq \inf_{z\in C^o} I(z)$. To this end, note that by [5, (3.5) and Theorem 3],

$$H(\Pi/P_X) = \inf_{\{Q \text{ of compact support}, \int x Q(dx) \in \overline{C}\}} H(Q|Q_X)$$

$$= \inf_{\{Q \text{ of compact support}, \int x Q(dx) \in C^o\}} H(Q|Q_X)\,. \tag{3.1}$$

Note that if $dQ/dQ_X = f$ is of compact support, then $\int |x| Q(dx) < \infty$ and for every $\lambda \in \mathbb{R}^\ell$,

$$\begin{aligned} H(Q|Q_X) &= \langle \lambda, \textstyle\int x Q(dx)\rangle - \int 1_{f>0} f \log(e^{\langle \lambda, x\rangle}/f) Q_X(dx) \\ &\geqq \langle \lambda, \textstyle\int x Q(dx)\rangle - \Lambda(\lambda)\,, \end{aligned}$$

implying that $H(Q|Q_X) \geqq I(\int x Q(dx))$. Consequently, using (3.1), $H(\Pi|P_X) \geqq \inf_{z\in C^o} I(z)$. $\square$

Proof of Corollary 2.11. $\Pi = \Pi'$ is closed with $\{\nu : \int U(x,y)\nu(dx)\nu(dy) < 1\} \subset \Pi^o$ (see [7, Lemma 7.3.12]). By (C-1) and the boundedness of U, Π is completely convex, with $H(\Pi|P_X) \leqq H(\Pi^o|P_X) < \infty$ by (C-3). Hence, by Sanov's theorem $P_{L_n}(\Pi') > 0$ for all n large enough. It was shown in [16] (see also [7, proof of Theorem 7.3.16]) that $P^* = \gamma_{\beta^*} \in \Pi$ where for all $\beta \geqq 0$, γ_β is of the form

$$\frac{d\gamma_\beta}{dP_X} = \exp(-\beta(U\gamma_\beta(x) - g(\beta)) + H(\gamma_\beta|P_X))$$

with

$$U\gamma_\beta(x) = \int U(x,y)\gamma_\beta(dy), \qquad g(\beta) = \int U(x,y)\gamma_\beta(dx)\gamma_\beta(dy)\,.$$

and $\beta^* = \inf\{\beta \geqq 0 : g(\beta) \leqq 1\}$. In particular [7, Lemma 7.3.14], $g(\beta^*) = 1$ and by (C-2), $\beta^* > 0$. Let $f(x) = \overline{f}(x)/\beta^* = 1 - U\gamma_{\beta^*}(x)$ and $\tilde{U}(x,y) = 1 - U(x,y) - f(x) - f(y)$. Define $Z_n = n^{-1/2}\Sigma_{i=1}^n f(X_i)$ and $Y_n = n^{-1}\Sigma_{i=1}^n \Sigma_{j=1}^n \tilde{U}(X_i,X_j)$. Note that $\{L_n \in \Pi'\} = \{2Z_n + n^{-1/2}Y_n \geqq 0\}$. Thus, for all $C > c > 0$,

$$\begin{aligned} &(P^*)^n\left(L_n \in \Pi', \sum_{i=1}^n \overline{f}(X_i) < C\beta^* \log n\right) \\ &\quad = (P^*)^n(-0.5n^{-1/2}Y_n \leqq Z_n < Cn^{-1/2}\log n) \\ &\quad \geqq (P^*)^n(cn^{-1/2}\log n \leqq Z_n < Cn^{-1/2}\log n) \\ &\quad\quad - (P^*)^n(Y_n \leqq -2c\log n)\,. \end{aligned} \tag{3.2}$$

Denoting hereafter expectations under $(P^*)^n = (\gamma_{\beta^*})^n$ by $E^*(\,\cdot\,)$, we see that $E^*(f(X_i)) = 0$, and for all $i \neq j$ also $E^*(\tilde{U}(X_i,X_j)|X_i) = E^*(\tilde{U}(X_j,X_i)|X_i) = 0$.

By the Berry–Esseen theorem (cf. [1, Theorem 12.4]), for all n large enough

$$(P^*)^n(cn^{-1/2}\log n \leqq Z_n < Cn^{-1/2}\log n) \geqq \frac{(C-c)\log n}{(M+1)\sqrt{2\pi n}} - O(n^{-1/2})\,. \tag{3.3}$$

Let $F_n(\,\cdot\,)$ denote the distribution function of Y_n. Then (see [12]),

$$\sup_x |F_n(x) - F_\infty(x)| \leqq O(n^{-1/2})$$

where F_∞ denotes the distribution function of the random variable $\theta = E^*(\tilde{U}(X,X)) + \Sigma_{j=1}^{\infty}\lambda_j(x_j^2 - 1)$, with λ_j deterministic, square summable, and x_j independent standard Normal random variables. It follows that

$$(P^*)^n(Y_n \leqq -2c\log n) \leqq P(\theta \leqq -2c\log n) + O(n^{-1/2})\,. \tag{3.4}$$

Due to the square summability of the λ_j there exists $\lambda_o > 0$ such that $c_1 = E(\exp(-\lambda_o\theta)) < \infty$. Hence, using Chebycheff's inequality,

$$P(\theta \leqq -2c\log n) \leqq c_1 e^{-2\lambda_o c\log n}\,. \tag{3.5}$$

Choose now $C > c > 1/(4\lambda_o)$, and combine (3.5) with (3.4) and (3.3) to conclude that (3.2) implies, for some $\eta > 0$ and all n large enough,

$$(P^*)^n\left(L_n \in \Pi', \sum_{i=1}^n \overline{f}(X_i) < C\beta^*\log n\right) \geqq \eta n^{-1/2}\log n\,.$$

The proof is completed by applying Proposition 2.8 for $\alpha_n = 0$ and $\rho_n = C\beta^* n^{-1}\log n$. $\square$

Proof of Proposition 2.12. Let $\sigma^2 = \int \overline{f}^2\,dP^*$ and $f = \overline{f}/\sigma$. Then $\int f\,dP^* = 0$ and $\int f^2\,dP^* = 1$, and, for any $v \in \Pi'$, $\langle f, v\rangle \geqq 0$ (see [5, (1.5)]). Let $T_n = \sqrt{k(n)}\langle f, L_{k(n)}\rangle$ and

$$V_n = \frac{1}{\beta}\sqrt{k(n)}\langle f, L_n\rangle = T_n + \frac{1}{\sqrt{k(n)}}\sum_{i=k(n)+1}^{n} f(X_i) = T_n + \sqrt{\frac{1-\beta}{\beta}}\,\overline{V}_n\,.$$

Denote the law of T_n (respectively, $\overline{V}_n$) under $(P^*)^n$ by $P_{n,T}$ (respectively, $P_{n,\overline{V}}$). Then, under our assumptions (cf. [1, Theorem 19.1]), $P_{n,T}$ and $P_{n,\overline{V}}$ possess densities denoted, respectively, by $p_{n,T}(t)$ and $p_{n,\overline{V}}(v)$, and, with $\phi(x) = (\sqrt{2\pi})^{-1}e^{-x^2/2}$,

$$\sup_{t\in\mathbb{R}} |p_{n,T}(t) - \phi(t)| = o(1)$$

$$\sup_{v\in\mathbb{R}} |p_{n,\overline{V}}(v) - \phi(v)| = o(1)\,.$$

Thus, it follows that the conditional law of T_n, conditioned on $V_n = v$, possesses a density $p_n(t|v)$, and, furthermore, denoting

$$\psi_\beta(t,v) = (\sqrt{2\pi(1-\beta)})^{-1}\exp[-(t-\beta v)^2/2(1-\beta)]\,,$$

one has that

$$\sup_{|t|,|v|\leqq 1} |p_n(t|v) - \psi_\beta(t,v)| = o(1)\,.$$

(The appearance of the density $\psi_\beta(t|v)$ is anything but mysterious: it represents the density of T conditioned on $T + \sqrt{(1-\beta)/\beta}\,\overline{V} = v$, where $T, \overline{V}$ are independent standard Normal).

Define now $F_n(v) = \exp[-\sigma\sqrt{k(n)}v](P^*)^n(L_n \in \Pi'|\langle f, L_n\rangle = \beta v/\sqrt{k(n)})$. Then $F_n(v) : \mathbb{R} \to [0,1]$ and $F_n(v) = 0$ for $v < 0$. Let $g(x) = 1_{\{-1\leqq x\leqq 1\}}$. Then, denoting by $P_n(v,t)$ the joint law of V_n, T_n under $(P^*)^n$,

$$\begin{aligned} E(g(T_n)|L_n \in \Pi') &= \frac{\int g(t)F_n(v)dP_n(v,t)}{\int F_n(v)dP_n(v,t)} \\ &= \int g(t)dQ_n(v,t) \end{aligned}$$

where

$$\frac{dQ_n(v,t)}{dP_n(v,t)} = \frac{F_n(v)}{\int F_n(v)dP_n(v)}$$

is independent of t. Note that, by (2.13),

$$\int F_n(v)dP_n(v,t) = P_{L_n}(\Pi')\exp[nH(P^*|P_X)] \geqq e^{-\mu(n)}\,.$$

Therefore, taking $\mu_1(n)\sqrt{n}/\max(\mu(n),1) \to_{n\to\infty} \infty$ but $\mu_1(n) = o(1)$, one has that

$$\int_{v\leqq\mu_1(n)} dQ_n(v,t) \geqq 1 - \exp[\mu(n) - \sigma\sqrt{k(n)}\mu_1(n)] \to_{n\to\infty} 1$$

implying that

$$\begin{aligned} \int_{\mathbb{R}} g(t)dQ_n(v,t) &\geqq \inf_{0\leqq v\leqq\mu_1(n)} E(g(T_n)|V_n = v) - o(1) \\ &= \inf_{0\leqq v\leqq\mu_1(n)} \int_{\mathbb{R}} g(t)p_n(t|v)dt - o(1) \\ &\geqq \inf_{0\leqq v\leqq\mu_1(n)} \int_{\mathbb{R}} g(t)\psi_\beta(t,v)dt - o(1) \\ &\to_{n\to\infty} \int_{\mathbb{R}} g(t)\psi_\beta(t,0)dt > \int_{\mathbb{R}} g(t)\phi(t)dt \end{aligned}$$

where the last inequality is due to the fact that $\psi_\beta(\,\cdot\,,0)$ is the density of Normal $(0, 1-\beta)$ law. On the other hand, by a standard CLT, $(P^*)^n(-1 \leqq T_n \leqq 1) \to_{n\to\infty} \int_{\mathbb{R}} g(t)\phi(t)dt$. One concludes that the variational distance between $P^n_{X^{k(n)}|\Pi'}$ and $(P^*)^{k(n)}$ is bounded away from zero, yielding, using [7, Exercise 6.2.17], the assertion (2.14). □

Proof of Proposition 2.15. By [5, (2.36)] $\overline{f}$ is an affine function of ψ and since $\{\lambda : \Lambda(\lambda) < \infty\}$ is an open set, all moments of $P^* \circ \psi^{-1}$ are finite. If the covariance matrix of $P^* \circ \psi^{-1}$ is singular, then for some $\lambda \in \mathbb{R}^\ell$ the random variable $\langle\lambda, \psi(X)\rangle$ is constant P_X-almost-surely. By removing all such deterministic relations from the definition of Π' we may and shall assume without loss of generality that this covariance matrix is positive definite. Hence, by an affine transformation of $\mathbb{R}^\ell$, we may assume hereafter that $\int \psi \, dP^* = 0$ and $\int \psi\psi' \, dP^*$ is the identity matrix. This transformation can be done such that $\overline{f} = \alpha\psi_1$ for some $\alpha > 0$, noting that then $C \subseteq \{v : v_1 \geqq 0\}$. Consequently, for any $A \subset \Sigma^k$ measurable,

$$P^n_{X_1^k|\Pi'}(A) = \int\limits_{A\times\Sigma^{n-k}} dP^n_{X_1^k|\Pi'}$$

$$= \frac{\int\limits_{A\times\Sigma^{n-k}} \mathbf{1}_{\{\frac{1}{n}\Sigma_{i=1}^n \psi(x_i)\in C\}} \exp\left[-\sum_{i=1}^n \alpha\psi_1(x_i)\right] d(P^*)^n}{\int\limits_{\Sigma^n} \mathbf{1}_{\{\frac{1}{n}\Sigma_{i=1}^n \psi(x_i)\in C\}} \exp\left[-\sum_{i=1}^n \alpha\psi_1(x_i)\right] d(P^*)^n}$$

Let $V_n = \frac{1}{\sqrt{n}}\Sigma_{i=1}^n \psi(X_i)$ for X_i i.i.d. P^* and $g_n(v) = \mathbf{1}_{\{v\in\sqrt{n}C\}} e^{-\sqrt{n}\alpha v_1}$. Denoting hereafter expectations under $(P^*)^n$ by $E^*(\,\cdot\,)$, we see that

$$\frac{dP_{X_1^k|\Pi'}}{dP^{*^k}} = h_n(V_k) = \frac{E^*[g_n(V_n)|V_k]}{E^*[g_n(V_n)]}$$

and

$$\|P^n_{X_1^k|\Pi'} - P^{*^k}\|_{\mathrm{var}} = E^*|h_n(V_k) - 1|\,.$$

Since $\|P^n_{X_1^k|\Pi'} - P^{*^k}\|_{\mathrm{var}}$ is monotone nondecreasing in k, it suffices to show that

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} E^*|h_n(V_{\varepsilon n}) - 1| = 0\,. \tag{3.6}$$

Under our assumptions, V_n possesses a bounded continuous density $p_n(v)$ which admits the asymptotic expansion

$$\sup_{v\in\mathbb{R}^\ell} (1 + \|v\|^3)|p_n(v) - \phi(v)(1 + n^{-1/2}H(v))| = o(n^{-1/2}) \tag{3.7}$$

where $\phi(v)$ is the standard Normal density on $\mathbb{R}^\ell$ and $H(v)$ is a polynomial of degree 3 in v (cf. [1, Theorem 19.2 and (7.20)]). The joint density of $(V_{\varepsilon n}, V_n)$ is then

$$p_n^\varepsilon(t, v) = p_{\varepsilon n}(t) p_{(1-\varepsilon)n}\left(\frac{v - \sqrt{\varepsilon}t}{\sqrt{1-\varepsilon}}\right)(1-\varepsilon)^{-\ell/2}\,.$$

With

$$b_n = \int\limits_{\mathbb{R}^\ell} g_n(v)p_n(v)dv = P_{L_n}(\Pi')\exp[nH(P^*|P_X)] \geqq M^{-1}n^{-1/2}$$

it follows that

$$E^*|h_n(V_{\varepsilon n}) - 1|$$

$$\leqq \frac{1}{b_n}\int\int g_n(v)|p_n^\varepsilon(t,v) - p_{\varepsilon n}(t)p_n(v)|dt\,dv \tag{3.8}$$

$$\leqq M\sqrt{n}\int\int g_n(v)p_{\varepsilon n}(t)\left|p_{(1-\varepsilon)n}\left(\frac{v-\sqrt{\varepsilon}t}{\sqrt{1-\varepsilon}}\right)(1-\varepsilon)^{-\ell/2} - p_n(v)\right| dt\,dv\,. \tag{3.9}$$

Due to the integrability of the error terms in (3.7), we may replace $p_k(\cdot)$ by $\phi(\cdot)(1+k^{-1/2}H(\cdot))$ when studying (3.9). Let $\overline{q}(\cdot)$ denote the centered Normal density for the covariance matrix $2I_\ell$. Note that $\phi(t)(1+|H(t)|) \leqq C\overline{q}(t)$ for some $C < \infty$ and all $t \in \mathbb{R}^\ell$. Moreover, differentiating with respect to $\sqrt{\varepsilon}$ it is straightforward to check that for all ε small enough,

$$\phi(t)\left|\phi\left(\frac{v-\sqrt{\varepsilon}t}{\sqrt{1-\varepsilon}}\right)(1-\varepsilon)^{-\ell/2}-\phi(v)\right| \leqq \sqrt{\varepsilon}C\overline{q}(t)\overline{q}(v),$$

for some $C < \infty$ and all $t, v \in \mathbb{R}^\ell$. Hence, for some $C_i < \infty$ independent of n and ε, and for all ε small enough and n large enough

$$\begin{aligned}
&\sqrt{n}\int\int g_n(v)p_{\varepsilon n}(t)\left|p_{(1-\varepsilon)n}\left(\frac{v-\sqrt{\varepsilon}t}{\sqrt{1-\varepsilon}}\right)(1-\varepsilon)^{-\ell/2}-p_n(v)\right|dt\,dv \\
&\leqq o(1)+C_1(\sqrt{n\varepsilon}+1)\int\int g_n(v)\overline{q}(t)\left[\overline{q}(v)+(1-\varepsilon)^{-\ell/2}\overline{q}\left(\frac{v-\sqrt{\varepsilon}t}{\sqrt{1-\varepsilon}}\right)\right]dt\,dv \\
&\leqq o(1)+C_2(\sqrt{n\varepsilon}+1)\int_0^\infty \exp[-\sqrt{n}\alpha v_1/2]\overline{q}(v_1)\,dv_1 \\
&\leqq o(1)+C_3(\varepsilon^{1/2}+n^{-1/2})
\end{aligned}$$

implying (3.6) and the proposition. □

Remark. 3.1. Let $d_n^\varepsilon(v) = \int |p_n^\varepsilon(t|v) - p_{\varepsilon n}(t)|dt \leqq 2$. For any compact $K \subseteq \mathbb{R}^\ell$ and all n large enough, $\inf_{v\in K} p_n(v) \geqq \frac{1}{2}\inf_{v\in K}\phi(v) > 0$. Therefore, by the same arguments as detailed before,

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\sup_{v\in K} d_n^\varepsilon(v) = 0\,.$$

If $q_n(v) = g_n(v)p_n(v)/b_n$ is a tight sequence in $M_1(\mathbb{R}^\ell)$ then

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\int q_n(v)d_n^\varepsilon(v)\,dv = 0\,.$$

By (3.8) we then get (3.6) and hence (2.17) even when (2.16) does not hold. The tightness of $\{q_n(\cdot)\}$ can be phrased in terms of the contact of C with $\int\psi\,dP^*$.

Acknowledgement. We thank P. Diaconis for several discussions concerning [8]. We also thank P. Bickel and A. Mandelbaum for guiding us through the U-statistics literature.

References

1. Bhattacharya, R.N., Ranga Rao, R.: Normal approximation and asymptotic expansions. New York: Wiley 1976
2. Bolthausen, E.: Laplace approximations for sums of independent random vectors, Part II. Probab. Theory Related Fields **76**, 167–206 (1987)

3. Bolthausen, E.: Maximum entropy principles for Markov processes. In: Stochastic Processes and Applications in Mathematical Physics, Proceeding of the Bielefeld Conference, 1985. Math. Appl. **61**, Dordrecht: Kluwer 1990
4. Csiszár, I.: I-divergence geometry of probability distributions and minimization problems. Ann. Probab. **3**, 146–158 (1975)
5. Csiszár, I.: Sanov property, generalized I-projection and a conditional limit theorem. Ann. Probab. **12**, 768–793 (1984)
6. Csiszár, I., Cover, T.M., Choi, B.S.: Conditional limit theorems under Markov conditioning. IEEE Trans. Inform. Theory **33**, 788–801 (1987)
7. Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications. Boston: Jones and Bartlett 1993
8. Diaconis, P., Freedman, D.: Finite exchangeable sequences. Ann. Probab. **4**, 745–764 (1980)
9. Diaconis, P., Freedman, D.: Conditional limit theorems for exponential families and finite versions of de Finetti's theorem. J. Theoret. Probab. **1**, 381–410 (1988)
10. Dinwoodie, I.H.: Measures dominantes et théorème de Sanov. Ann. Inst. Henry Poincaré **28**, 365–373 (1992)
11. Iltis, M.: Asymptotics of Large Deviations in $\mathbb{R}^d$. Preprint, 1993
12. Korolyuk, V.S., Borovskikh, Yu.V.: Convergence rate for degenerate Von Mises functionals. Theory Probab. Appl. **33**, 125–135 (1988)
13. Lanford, O.E.: Entropy and equilibrium states in classical statistical mechanics. Lecture notes in Physics Vol. 20, 1–113 (1973) New York: Springer
14. Ney, P.: Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$. Ann. Probab. **11**, 158–167 (1983)
15. Schroeder, C.: I-projection and conditional limit theorems for discrete parameter Markov processes. Ann. Probab. **21**, 721–758 (1993)
16. Stroock, D.W., Zeitouni, O.: Microcanonical distributions, Gibbs states, and the equivalence of ensembles. In: Durrett, R., Kesten, H. (eds.) Festchrift in honor of F. Spitzer, pp. 399–424. Basel, Switzerland: Birkhäuser, 1991
17. von Bahr, B., Esseen, C.G.: Inequalities for the rth absolute moment for a sum of random variables, $1 \leqq r \leqq 2$. Ann. Math. Statistics **36**, 299–303 (1965)