Refinements of the Gibbs conditioning principle

Amir Dembo *

Department of Mathematics and Department of Statistics

Stanford University

Ofer Zeitouni †

Department of Electrical Engineering

Technion- Israel Institute of Technology, Haifa, ISRAEL

April 10, 1994. Revised: May 8, 1995

Abstract 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. Extensions of Csiszár's triangle inequality and information theoretic identity to the Markov chain set–up lead to similar refinements in the Markov case.

1 Introduction

Throughout this paper, $X_1, X_2, \ldots$ denotes a sequence of independent, identically distributed random variables, distributed over a Polish space $(\Sigma, \mathcal{B}_\Sigma)$ with common distribution P_X. Here, $\mathcal{B}_\Sigma$ denotes the Borel σ-field of Σ. Let $L_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denote the empirical measure of the sequence

*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.

Key Words: Gibbs conditioning, large deviations, empirical measure, Markov chains.

AMS (1991) subject classifications: 60F10

$\{X_i\}_{i=1}^n$, 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 Section 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 $A(a,\delta) = \{\omega : n^{-1}\sum_{i=1}^n f(X_i) \in [a-\delta, a+\delta]\}$, for some measurable function $f(\cdot)$. Under suitable conditions on P_X and $f(\cdot)$, the Gibbs conditioning principle is the statement that, for any Borel set $B \in \mathcal{B}_\Sigma$, 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, 18, 23]).

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 that 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 [23]. 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 $C \in \mathcal{B}_\Sigma^k$. One may also obtain Markov analogues of these results (see, e.g., [3, 6, 22, 23]).

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 ([22]).

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 Section 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), and, at the price of less rapid growth of $k(n)$, under conditioning by an infinite class of functions (see Proposition 2.18 and the examples following it). These results form the core of Section 2. For the sake of better readability, we have postponed many proofs which interrupt the flow of the presentation in Section 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.

Having dealt with the i.i.d. case, we turn our attention in Section 4 to the Markov situation. Extensions of the basic results of Csiszár to the Markov set-up have been recently derived in [6] (for Σ a finite set) by a counting approach, and in [22] (for Σ a compact metric space) by using instead the large deviations results of Donsker and Varadhan. We present here an alternative derivation, which is closer in spirit to the original approach of Csiszár. Indeed, in the i.i.d. case, the latter was based on the geometric observation (the "triangle inequality" of [4, (2.14)]) that, under suitable technical conditions, if A is a closed convex set of probability measures, $R \notin A$ and $Q \in A$ satisfies $\inf_{\nu \in A} H(\nu|R) = H(Q|R)$, then, for any $P \in A$,

$$H(P|R) \geq H(Q|R) + H(P|Q) .$$

(the geometrical picture being that if $H(\cdot|\cdot)$ is interpreted as the square of the distance, then the angle between the lines connecting R, Q and Q, P is acute). In the Markov case, we prove a similar statement, where the main objects of study are transition kernels (see Lemma 4.5). This geometric observation, coupled with the Markov chain duals of the information identity of [5, (2.11)] (see (4.2)), allows us to repeat the analysis of the i.i.d. refinements in the Markov case. In particular, a corollary of our results (see Corollary 4.17) is that in the Markov case, the convergence in variation norm to an appropriate γ^* as in (1.2) holds true under simple energy conditioning as soon as $k(n) \log n/n \to_{n\to\infty} 0$.

We conclude this introduction with some 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 uses Proposition 2.18). Next, the Markov chain results 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 $(\Omega, \mathcal{B}, \mu)$ and Markov kernel ν from $(\Omega, \mathcal{B})$ to $(\Sigma, \mathcal{B}_\Sigma)$ such that $\nu(\omega, \cdot) \in \Pi$ for each $\omega \in \Omega$, the probability measure $\mu\nu$ defined by $\mu\nu(\cdot) = \int \nu(x, \cdot)\mu(dx)$ 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 $\nu \in \Pi$ 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 $L_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ 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* $H(\Pi|P_X) = \inf_{P\in\Pi} H(P|P_X) < \infty$.

Here, Π' measurable means that $\{(x_1, \ldots, x_n) : n^{-1}\sum_{i=1}^n \delta_{x_i} \in \Pi'\} \in \mathcal{B}_{\Sigma^n} (= (\mathcal{B}_\Sigma)^n)$ 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,*

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

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 i-th coordinate (see [5, (2.10)]),

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

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

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

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 j-th 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

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

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

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

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 $\{n^{-1}\sum_{i=1}^n \psi(X_i) \in C\}$, 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 $\Pi' = \{\nu : \nu \circ \psi^{-1}$ of compact support, $\int \psi d\nu \in C\}$, 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 $\int x Q_X(dx) \notin C$. 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 Section 3. ∎

The rate in Corollary 2.7 is in general not optimal. As will be shown below (c.f. 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 $n^{-1}\sum_{i=1}^n X_i \geq 1$. 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 for this lies in (2.5) where we ignored the contribution due to the conditional dependence among the $k(n)$-blocks.

In [10, Theorem 1.6], Diaconis and Freedman deal with point conditioning, as in the above example when taking $C = \{1\}$, 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 [10, Lemma 3.1] suffices for convergence in divergence). In the setting of [10], $P^n_{X^n|\Pi'} = (P^*)^n_{X^n|\Pi'}$ 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 $H((P^*)^n_{X^n|\Pi'}|(P^*)^n) \le \log 2$ 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 [9, Theorem 13] that the variational distance between $P^n_{X^k|\Pi'}$ 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 [9, Proposition 31] give an example of Π' for which this rate is tight. In comparison, our results deal with the 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.

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*

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

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

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

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

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

(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 $\rho > 0$

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

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 [12, 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) $\int U(x,y)(Q_1 - Q_2)(dx)(Q_1 - Q_2)(dy) \geq 0$ 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, and let* $\Pi = \Pi' = \{\nu : \int U(x,y)\nu(dx)\nu(dy) \leq 1\}$. *Then,* $H(P^n_{X^{k(n)}|\Pi'} | (P^*)^{k(n)}) \to 0$ *provided that* $n^{-1} k(n) \log n \to 0$.

Remark: Note that the conditioning $L_n \in \Pi'$ corresponds to $n^{-2} \sum_{i,j=1}^n U(X_i, X_j) \leq 1$.

Proof: See Section 3. ■

Having spent some effort in obtaining convergence statements for the conditional law $P^n_{X^{k(n)}|\Pi'}$, and before turning to the infinite dimensional setup, 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*

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

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 Section 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

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* $M < \infty$,

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

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 Section 3. ■

Remarks: (1) Conditions for (2.16) to hold are given in [15, 20]. 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) (c.f. Remark 3.1).

Proof: See Section 3. ■

We next turn to the infinite dimensional setup. Let $\mathcal{F}$ be a permissible class (see [21, Appendix C]) of Borel measurable functions from Σ to $[-1,1]$. Equip the set $\mathcal{X}$ of bounded real valued functions on $\mathcal{F}$ with the norm $\|x\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|x(f)|$. We shall identify $M(\Sigma)$ with a subset of $\mathcal{X}$ via the mapping $\nu \mapsto \nu_{\mathcal{F}}$ such that $\nu_{\mathcal{F}}(f) = \int f d\nu$. Denote by $B_{\mu,\delta} = \{\nu \in M_1(\Sigma) : \|\nu - \mu\|_{\mathcal{F}} < \delta\}$ the $\mathcal{F}$-ball of radius δ, centered at $\mu \in M_1(\Sigma)$.

Let $N_d(\delta, \mathcal{F})$ denote the minimal cardinality of a δ-cover of $\mathcal{F}$ in the (pseudo) metric space $(B(\Sigma), d)$, and for any $\mathbf{x}_n = (x_1, \ldots, x_n) \in \Sigma^n$ define the (pseudo) metric $\ell^1_{\mathbf{x}_n}(f,g) = n^{-1}\sum_{i=1}^n |f(x_i) - g(x_i)|$ on $B(\Sigma)$. Finally, let $H_n(\delta, \mathcal{F}) = \sup_{\mathbf{x}_n\in\Sigma^n} \log N_{\ell^1_{\mathbf{x}_n}}(\delta, \mathcal{F})$, and recall that $\mathcal{F}$ is a uniform Glivenko-Cantelli class of functions iff $H_n(\delta,\mathcal{F})/n \to 0$ for any fixed $\delta > 0$ (see [13, Theorem 6]).

Proposition 2.18 *Assume (A-1) and that* $\int (\log \frac{dP^*}{dP_X})^2 dP^* < \infty$. *Suppose that for some* $Q \in \Pi'$ *with* $\int(\log\frac{dQ}{dP_X})^2 dQ < \infty$, *it holds that* $B_{Q_{\alpha_n}, 8\delta_n} \subset \Pi'$ *for some* $\alpha_n \geq \delta_n \to 0$ *such that* $n\delta_n^2 - 2H_n(\delta_n, \mathcal{F}) \to \infty$. *Then,* $H(P^n_{X^{k(n)}|\Pi'} | (P^*)^{k(n)}) \to 0$ *for any* $k(n)$ *such that* $k(n)\alpha_n \to 0$.

Proof: See Section 3. ■

Remarks: 1) Note that in Proposition 2.18, $k(n)$ is at most of $o(\sqrt{n})$, demonstrating a significant loss compared with the finite dimensional situation.
2) In Proposition 2.18 the conditions $\overline{f} \in L_2(P^*)$ and $\log \frac{dQ}{dP_X} \in L_2(Q)$ may be relaxed to $\overline{f} \in L_\beta(P^*)$ and $\log\frac{dQ}{dP_X} \in L_\beta(Q)$ for some $\beta \in (1,2)$ at the expense of requiring $n^{1-1/\beta}\delta_n$ to be bounded away from zero (c.f. Remark 3.2).

In order to avoid discussing measurability issues, we restrict the discussion in the next proposition to the case of countable $\mathcal{F}$. Instead, one could put a separability condition on the empirical process indexed by $\mathcal{F}$. Recall that $\mathcal{F}$ is a Donsker class with respect to P^* iff for X_i i.i.d. P^* random variables, the laws of $V_n = \sqrt{n}(L_n - P^*)_{\mathcal{F}}$ converge in the $C_b(\mathcal{X})$-topology of $M_1(\mathcal{X})$ to a Gaussian Radon measure γ_{P^*} on $\mathcal{X}$ (c.f. [19, Section 14.2]). This suggests the following alternative to Proposition 2.18.

Proposition 2.19 *Assume (A-1) and that* $\int(\log\frac{dP^*}{dP_X})^2 dP^* < \infty$. *Suppose there exists a Donsker class* $\mathcal{F}$ *with respect to* P^* *and* $C \subset \mathcal{X}$ *open, convex such that* $P^*_{\mathcal{F}} \in \overline{C}$ *and* $\{\nu \in M_1(\Sigma) : \nu_{\mathcal{F}} \in$

$C\} \subset \Pi'$. *If* $\gamma_{P^*}(\{t(x - P^*_{\mathcal{F}}) : x \in C\}) > 0$ *for some* $t \in (0, \infty)$ *then* $H(P^n_{X^{k(n)}|\Pi'} | (P^*)^{k(n)}) \to 0$ *for any* $k(n) = o(\sqrt{n})$.

Proof: See Section 3. ∎

The following lemma, adapted from [12, Theorem 2.1], is a useful tool in the application of Propositions 2.18 and 2.19.

Lemma 2.20 *Let* $M_1(\Sigma)$ *be equipped with the* V*-topology, where either* $V = C_b(\Sigma)$ *(the weak topology) or* $V = B(\Sigma)$ *(the* τ*-topology). Let* Π *be convex, open and such that* $H(\Pi|P_X) < \infty$, *and* $P^* \in \overline{\Pi}$ *denote the generalized I–projection of* P_X *on* Π. *Then,* $\log(\frac{dP^*}{dP_X}) \in V$.

Proof: See Section 3. ∎

We conclude this section with two examples demonstrating the applicability of Proposition 2.18 and Lemma 2.20.

Example 2.1 Let Lip(1,1) denote the class of Lipschitz continuous functions, bounded in absolute value by 1, with Lipschitz constant bounded by 1. Let $\mathcal{F}$ consist of a permissible class of Lip(1,1) functions. Let $Q \in M_1(\Sigma)$ be such that $\|P_X - Q\|_{\mathcal{F}} > \gamma$ for some $\gamma \in (0, 2)$, and such that $\log(dQ/dP_X) \in L_2(Q)$. By Fubini's theorem, $\Pi' = \Pi = \{\nu : \|\nu - Q\|_{\mathcal{F}} < \gamma\}$ is completely convex. With respect to the $C_b(\Sigma)$-topology on $M_1(\Sigma)$, the set Π is open with $H(\Pi|P_X) \leq H(Q|P_X) < \infty$. Thus, (A–1) is satisfied since by Sanov's theorem $P_{L_n}(\Pi') \geq e^{-n(H(Q|P_X)+1)} > 0$ for n large enough. Moreover, by Lemma 2.20, $P^* \in \overline{\Pi}$ and $\log(\frac{dP^*}{dP_X}) \in C_b(\Sigma)$. Next, note that for every $\nu \in M_1(\Sigma)$

$$\|\nu - Q\|_{\mathcal{F}} \leq \|\nu - Q_\alpha\|_{\mathcal{F}} + \|Q_\alpha - Q\|_{\mathcal{F}} \leq \|\nu - Q_\alpha\|_{\mathcal{F}} + (1 - \alpha)\gamma \ ,$$

implying that $B_{Q_\alpha, 8\delta} \subset \Pi'$ for all $\alpha \geq 8\delta/\gamma$. Choosing $\alpha_n = 8\delta_n/\gamma$, δ_n such that $n\delta_n^2 - 2H_n(\delta_n, \mathcal{F}) \to \infty$ and $k(n) = o(\delta_n^{-1})$ we conclude by Proposition 2.18 that $H(P^n_{X^{k(n)}|\Pi'} | (P^*)^{k(n)}) \to 0$. ∎

The next example demonstrates the additional work needed in order to apply Proposition 2.18 when the conditioning set possesses no interior.

Example 2.2 Let $\Sigma = \mathbb{R}$ and $\mathcal{F} = \mathcal{F}_I = \{1_{(b,c]}(x) : -\infty \leq b < c \leq \infty\}$. Let Q and P_X be such that $\|Q - P_X\|_{\mathcal{F}} > \gamma$, some $\gamma \in (0, 1)$ and $\frac{dQ}{dP_X} = \sum_{i=1}^{L} \kappa_i f_i$, $L < \infty$ and $f_i = 1_{(a_i, a_{i+1}]} \in \mathcal{F}_I$ with

$-\infty = a_1 < \cdots < a_{L+1} = \infty$. Clearly, $\int \log(\frac{dQ}{dP_X})^2 dQ < \infty$. Consider the variation-closed set $\Pi' = \Pi = \{\nu : \|\nu - Q\|_{\mathcal{F}} \leq \gamma\}$. Note that $L_n \in \Pi'$ is closely related to conditioning on the empirical measure being in a Kolmogorov-Smirnov ball centered at Q. Clearly, $P_X \notin \Pi$ while $P^* \in \Pi$. Use the change of measure as in the first equality of (2.10) together with the boundedness of $\frac{dQ}{dP_X}$ to conclude that $P_{L_n}(\Pi') \geq c^n Q_{L_n}(\Pi')$ for some $c > 0$. By the classical Glivenko-Cantelli theorem (see [21, Section II.3]) $Q_{L_n}(\Pi') \to 1$ as $n \to \infty$, leading to the conclusion that $P_{L_n}(\Pi') > 0$ for all n large. Since, by Fubini's theorem, Π is completely convex, (A-1) follows. We show below that

$$\frac{dP^*}{dP_X} = \sum_{i=1}^{L} \xi_i f_i \text{ for some } \xi_1, \ldots, \xi_L \, , \tag{2.21}$$

implying that $\int (\log \frac{dP^*}{dP_X})^2 dP^* < \infty$.

Note that, exactly as in Example 2.1, $B_{Q_\alpha, 8\delta} \subset \Pi'$ for all $\alpha \geq 8\delta/\gamma$. Since $H_n(\delta, \mathcal{F}_I) \leq 2\log(n+1)$ for all $\delta > 0$. It follows that for c large enough the choice $\delta_n = c\sqrt{\log n/n}$ with $\alpha_n = 8\delta_n/\gamma$ and $k_n = o(\sqrt{n/\log n})$ allows for the application of Proposition 2.18.

We turn finally to showing (2.21). Consider the auxiliary problem,

$$\inf_{\{P : |\int f dP - \int f dQ| \leq \gamma, f \in \mathcal{F}_L\}} H(P|P_X)$$

where $\mathcal{F}_L = \{\sum_{i=j}^{k} f_i : 1 \leq j \leq k \leq L\}$, whose solution, denoted $\hat{P}$ is easily checked to be of the form (2.21). Since $\mathcal{F}_L \subset \mathcal{F}_I$, suffices to check that $\hat{P} \in \Pi$ in order to conclude that $P^* = \hat{P}$. To this end, fix $f = 1_{(b,c]} \in \mathcal{F}_I$ such that $b \in [a_j, a_{j+1}]$ and $c \in [a_k, a_{k+1}]$. Assume first that $\hat{P}((b,c]) - Q((b,c]) > 0$. Since the measure $\hat{P} - Q$ has a fixed sign on $[a_j, a_{j+1}]$, either $\hat{P}((b,c]) - Q((b,c]) \leq \hat{P}((a_j,c]) - Q((a_j,c])$ or $\hat{P}((b,c]) - Q((b,c]) \leq \hat{P}((a_{j+1},c]) - Q((a_{j+1},c])$. Similar arguments apply to $[a_k, a_{k+1}]$ and to the case of $\hat{P}((b,c]) - Q((b,c]) < 0$, leading to the conclusion that

$$|\hat{P}((b,c]) - Q((b,c])| \leq \sup_{f \in \mathcal{F}_L} |\int f d\hat{P} - \int f dQ| \leq \gamma \, .$$

This example easily generalizes to $\mathcal{F}_I$ consisting of indicators on rectangles in $\mathbb{R}^d$. The generalization to measures Q of general structure and to other Glivenko-Cantelli classes is similar, except for the bound $\log \frac{dP^*}{dP_X} \in L_\beta(P^*)$, $\beta > 1$, which has to be provided on a case by case basis. ■

3 Proofs for Section 2

Proof of Corollary 2.7 Let $I(z) = \sup_{\lambda\in\mathbb{R}^\ell}[\langle\lambda, z\rangle - \Lambda(\lambda)]$. Note that $P_{L_n}(\Pi') = Q_X^n(n^{-1}\sum_{i=1}^n Y_i \in C)$, where $Y_i = \psi(X_i)$ are i.i.d. Q_X. It follows from [20, (3.4)] that for some finite $c_1 > 0$ and n large enough

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

where

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

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 $\Pi = \{\nu : \nu\circ\psi^{-1} \text{ of compact support}, \int\psi d\nu \in \overline{C}\}$, 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 $H(\Pi|P_X) \geq \inf_{z\in C^o} I(z)$. To this end, note that by [5, (3.5) and Theorem 3],

$$\begin{aligned} H(\Pi|P_X) &= \inf_{\{Q \text{ of compact support}, \int xQ(dx)\in\overline{C}\}} H(Q|Q_X) \\ &= \inf_{\{Q \text{ of compact support}, \int xQ(dx)\in C^o\}} H(Q|Q_X) . \end{aligned} \tag{3.1}$$

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

$$H(Q|Q_X) = \langle\lambda, \int xQ(dx)\rangle - \int 1_{f>0} f\log(e^{\langle\lambda,x\rangle}/f)Q_X(dx) \geq \langle\lambda, \int xQ(dx)\rangle - \Lambda(\lambda) ,$$

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

Proof of Corollary 2.11 $\Pi = \Pi'$ is closed with $\{\nu : \int U(x,y)\nu(dx)\nu(dy) < 1\} \subset \Pi^o$ (see [8, Lemma 7.3.12]). By (C–1) and the boundedness of U, Π is completely convex, with $H(\Pi|P_X) \leq H(\Pi^o|P_X) < \infty$ by (C–3). Hence, by Sanov's theorem $P_{L_n}(\Pi') > 0$ for all n large enough. It was shown in [23] (see also [8, proof of Theorem 7.3.16]) that $P^* = \gamma_{\beta^*} \in \Pi$ where for all $\beta \geq 0$, 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

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

and $\beta^* = \inf\{\beta \geq 0 : g(\beta) \leq 1\}$. In particular ([8, 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 $Z_n = n^{-1/2}\sum_{i=1}^n f(X_i)$ and $Y_n = n^{-1}\sum_{i=1}^n \sum_{j=1}^n \tilde{U}(X_i, X_j)$. Note that $\{L_n \in \Pi'\} = \{2Z_n + n^{-1/2}Y_n \geq 0\}$. Thus, for all $C > c > 0$,

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

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 (c.f. [1, Theorem 12.4]), for all n large enough

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

Let $F_n(\cdot)$ denote the distribution function of Y_n. Then (see [17]),

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

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

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

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,

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

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

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

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

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 $\nu \in \Pi'$, $\langle f, \nu \rangle \geq 0$ (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 $P_{n,T}$ ($P_{n,\overline{V}}$). Then, under our assumptions (c.f. [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 $\phi(x) = (\sqrt{2\pi})^{-1} e^{-x^2/2}$,

$$\begin{aligned} \sup_{t \in \mathbb{R}} |p_{n,T}(t) - \phi(t)| &= o(1) \\ \sup_{v \in \mathbb{R}} |p_{n,\overline{V}}(v) - \phi(v)| &= o(1) \, . \end{aligned}$$

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

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

one has that

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

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

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

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),

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

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

$$\int_{v\le\mu_1(n)} dQ_n(v,t) \ge 1 - e^{\mu(n)-\sigma\sqrt{k(n)}\mu_1(n)} \to_{n\to\infty} 1\,,$$

implying that

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

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, $(P^*)^n(-1\le T_n\le 1) \to_{n\to\infty} \int_{\mathbb{R}} g(t)\phi(t)dt$. One concludes that the variational distance between $P^n_{X^{k(n)}|\Pi'}$ and $(P^*)^{k(n)}$ is bounded away from zero, yielding, using [8, 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 $C\subseteq\{v : v_1\ge 0\}$. Consequently, for any $A\subset\Sigma^k$ measurable,

$$P^n_{X_1^k|\Pi'}(A) = \int_{A\times\Sigma^{n-k}} dP^n_{X_1^n|\Pi'} = \frac{\int_{A\times\Sigma^{n-k}} \mathbf{1}_{\{\frac{1}{n}\sum_{i=1}^n \psi(x_i)\in C\}} e^{-\sum_{i=1}^n \alpha\psi_1(x_i)}\, d(P^*)^n}{\int_{\Sigma^n} \mathbf{1}_{\{\frac{1}{n}\sum_{i=1}^n \psi(x_i)\in C\}} e^{-\sum_{i=1}^n \alpha\psi_1(x_i)}\, d(P^*)^n}$$

Let $V_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i)$ 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

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

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

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

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 (c.f. [1, Theorem 19.2 and (7.20)]). The joint density of $(V_{\epsilon n}, V_n)$ is then

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

With

$$b_n = \int_{\mathbb{R}^\ell} g_n(v)p_n(v)dv = P_{L_n}(\Pi')e^{nH(P^*|P_X)} \geq M^{-1}n^{-1/2} \,,$$

it follows that

$$E^*|h_n(V_{\epsilon n}) - 1| \leq \frac{1}{b_n}\iint g_n(v)|p_n^\epsilon(t,v) - p_{\epsilon n}(t)p_n(v)|dt\,dv \tag{3.8}$$

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

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 $\phi(t)(1+|H(t)|) \leq C\overline{q}(t)$ for some $C<\infty$ and all $t\in\mathbb{R}^\ell$. Moreover, differentiating with respect to $\sqrt{\epsilon}$ it is straightforward to check that for all ϵ small enough,

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

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

$$\sqrt{n}\iint g_n(v)p_{\epsilon n}(t)|p_{(1-\epsilon)n}\left(\frac{v-\sqrt{\epsilon}t}{\sqrt{1-\epsilon}}\right)(1-\epsilon)^{-\ell/2} - p_n(v)|dt\,dv$$

$$
\begin{aligned}
&\leq o(1) \quad +C_1(\sqrt{n\epsilon}+1)\iint g_n(v)\overline{q}(t)[\overline{q}(v)+(1-\epsilon)^{-\ell/2}\overline{q}\left(\frac{v-\sqrt{\epsilon}t}{\sqrt{1-\epsilon}}\right)]dt\,dv \\
&\leq o(1) \quad +C_2(\sqrt{n\epsilon}+1)\int_0^\infty e^{-\sqrt{n}\alpha v_1/2}\overline{q}(v_1)dv_1 \\
&\leq o(1) \quad +C_3(\epsilon^{1/2}+n^{-1/2}),
\end{aligned}
$$

implying (3.6) and the proposition. ■

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

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

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

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

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 $\int\psi dP^*$.

Proof of Proposition 2.18 Let f_α be as in Proposition 2.8. Let $1\leq\beta\leq 2$. Since $h(x)=x|\log x|^\beta$ is bounded on $[0,1]$ by $h(e^{-\beta})<\infty$, and $h(x)1_{\{x\geq 1\}}$ is convex, it follows that for some constant $M<\infty$,

$$
\begin{aligned}
\int|f_\alpha|^\beta dQ_\alpha &= \int h(\alpha\frac{dQ}{dP_X}+(1-\alpha)\frac{dP^*}{dP_X})dP_X \\
&\leq h(e^{-\beta})+\alpha\int|\log\frac{dQ}{dP_X}|^\beta dQ+(1-\alpha)\int|\log\frac{dP^*}{dP_X}|^\beta dP^*\leq M\,. \qquad (3.10)
\end{aligned}
$$

Hence, by Chebycheff's inequality (with $\beta=2$),

$$e_n^{(1)}=(Q_{\alpha_n})^n(n^{-1}|\sum_{i=1}^n(f_{\alpha_n}(X_i)-\int f_{\alpha_n}dQ_{\alpha_n})|\geq\delta_n)\leq\frac{M}{\delta_n^2 n}\to 0\,. \qquad (3.11)$$

Moreover, for n large enough such that $n\delta_n^2\geq 1/8$, we have by [21, page 31, inequalities (30),(31)] that

$$e_n^{(2)}=(Q_{\alpha_n})^n(L_n\notin\Pi')\leq(Q_{\alpha_n})^n(L_n\notin B_{Q_{\alpha_n},8\delta_n})\leq 8e^{H_n(\delta_n,\mathcal{F})-n\delta_n^2/2}\to 0\,.$$

(The term $(Q_{\alpha_n})^n(L_n\notin B_{Q_{\alpha_n},8\delta_n})$ is well defined since $\mathcal{F}$ is permissible). In view of Proposition 2.8, suffices to prove that (2.9) holds for $\rho_n=\delta_n$. This follows by the union of events bound since $e_n^{(1)}+e_n^{(2)}\to 0$. ■

Remark 3.2 Proposition 2.18 remains valid if the conditions $\int (\log \frac{dP^*}{dP_X})^2 dP^* < \infty$ and $\int (\log \frac{dQ}{dP_X})^2 dQ < \infty$ are replaced by $\int |\log \frac{dP^*}{dP_X}|^\beta dP^* < \infty$ and $\int |\log \frac{dQ}{dP_X}|^\beta dQ < \infty$, respectively, for some $1 < \beta < 2$, as soon as the other conditions hold for $\delta_n = cn^{1/\beta - 1}$ and c large. Indeed, $e_n^{(2)}$ is handled exactly as before, while one has

$$e_n^{(1)} \leq \frac{2nE_{Q_{\alpha_n}}(|f_{\alpha_n}(X_1) - \int f_{\alpha_n} dQ_{\alpha_n}|^\beta)}{(\delta_n n)^\beta} \leq 22^\beta M n^{1-\beta} \delta_n^{-\beta} = 2M(2/c)^\beta < 1,$$

where the first inequality is due to Chebycheff's inequality (for x^β) combined with [24, Theorem 2], and the second follows from (3.10).

Proof of Proposition 2.19 Let $e_n^{(1)} = (P^*)^n(n^{-1}|\sum_{i=1}^n \overline{f}(X_i)| \geq cn^{-1/2})$ and $e_n^{(2)} = (P^*)^n(L_n \notin \Pi')$. Since $e_n^{(1)} \leq M/c^2$ (c.f. (3.11)), by Proposition 2.8 suffices to show that $\limsup_{n\to\infty} e_n^{(2)} < 1$ (c.f. (2.9) for $\alpha_n = 0$ and $\rho_n = cn^{-1/2}$). (We note that measurability questions can be taken into account by using outer measures when bounding $e_n^{(2)}$.) Let $V_n = \sqrt{n}(L_n - P^*)_{\mathcal{F}}$ and $T_s = \{s(x - P^*_{\mathcal{F}}) : x \in C\}$ which are open subsets of $\mathcal{X}$ for $s \in (0, \infty)$. By our assumptions $e_n^{(2)} \leq 1 - (P^*)^n((L_n)_{\mathcal{F}} \in C) = 1 - (P^*)^n(V_n \in T_{\sqrt{n}})$. Since C is an open convex set and $P^*_{\mathcal{F}} \in \overline{C}$, it follows that T_s is non-decreasing in s. Hence, $e_n^{(2)} \leq 1 - (P^*)^n(V_n \in T_t)$ for all $n \geq t^2$. The $C_b(\mathcal{X})$-convergence of the laws of V_n to γ_{P^*} thus implies that $\limsup_{n\to\infty} e_n^{(2)} \leq 1 - \gamma_{P^*}(T_t) < 1$. ■

Proof of Lemma 2.20 We may and will assume that $P^* \neq P_X$ for otherwise there is nothing to prove. In particular, $H(\Pi|P_X) > 0$. With the V-topology weaker than the variational norm topology on $M_1(\Sigma)$, clearly $P^* \in \overline{\Pi}$. Consider the convex, compact set

$$L = \{\nu \in M_1(\Sigma) : H(\nu|P_X) \leq H(\Pi|P_X)\}.$$

Since Π is open, if $P \in L \cap \Pi$ then $P(t) = (1-t)P + tP_X \in \Pi$ for $t > 0$ small enough, with $H(P(t)|P_X) \leq (1-t)H(P|P_X) < H(\Pi|P_X)$. Therefore, the convex, open set Π and L are disjoint subsets of $M_1(\Sigma)$, and by the Hahn–Banach theorem (cf. [11, Theorem V.2.8]), there exists $h \in M(\Sigma)^* = V$ and $\alpha \in \mathbb{R}$ such that for $g = h - \alpha \in V$

$$L \subset \{\nu : \langle g, \nu\rangle \leq 0\}, \quad \Pi \subset \{\mu : \langle g, \mu\rangle > 0\}.$$

In particular,

$$H(\Pi|P_X) = \inf_{\{\nu : \langle g,\nu\rangle > 0\}} H(\nu|P_X) = \inf_{x>0} \sup_{t\in\mathbb{R}} \{tx - \log E_{P_X}(e^{tg})\} \tag{3.12}$$

(since $\nu \mapsto \langle g, \nu \rangle$ is a continuous mapping, the rightmost equality follows for example by comparing [8, Theorem 2.2.3] with [8, Theorem 6.2.10]).

Since $P_X \in L$ and there exists $Q \ll P_X$, $Q \in \Pi$ it follows that $\langle g, P_X \rangle \leq 0$ and moreover, $P_X(\{g > 0\}) > 0$ and $P_X(\{g < 0\}) > 0$. With g bounded, hence $E_{P_X}(ge^{t_0 g}) = 0$ for some $\infty > t_0 \geq 0$ and

$$\begin{aligned} \inf_{x>0} \sup_{t \in \mathbb{R}} \{tx - \log E_{P_X}(e^{tg})\} &= \lim_{x \downarrow 0} \sup_{t \geq 0} \{tx - \log E_{P_X}(e^{tg})\} \\ = \sup_{t \geq 0} \{-\log E_{P_X}(e^{tg})\} &= -\log E_{P_X}(e^{t_0 g}) \end{aligned} \tag{3.13}$$

(c.f. [8, Lemma 2.2.5]).

Let $f = t_0 g$. Combining (3.12) and (3.13) we see that $H(\Pi | P_X) = -\log E_{P_X}(e^f)$ and $\int f dP \geq 0$ for every $P \in \Pi$. Therefore, by [5, Lemma 3.1], $\log(\frac{dP^*}{dP_X}) = f + H(\Pi | P_X) \in V$. ■

4 Refinements in the Markov case

Hereafter, we consider the measure space $(\Omega, \mathcal{B}_\Omega)$ with $\Omega = \Sigma^{\mathbb{Z}_+}$ and Σ a Polish space. The marginal of $Q \in M_1(\Omega)$ on the coordinates $(i, i+1, \ldots, j)$, $j \geq i$ is denoted $Q_{i,j}$ (or Q_i when $j = i$), and for $j > i$ the corresponding regular conditional probability distributions are denoted $Q_{i+1,j|i}(x, \cdot)$, (or $Q_{i+1|i}(x, \cdot)$ when $j = i+1$). The Markov measure on Σ^{k+1} corresponding to initial distribution ν and transition kernel π is denoted $\nu \otimes_k \pi$, omitting the index k for $k = 1$. Similarly, $\nu \otimes \pi_1 \otimes \pi_2 \otimes \cdots \otimes \pi_k$ denotes the non-homogeneous Markov measure induced by the kernels $\pi_1, \ldots, \pi_k$. If $Q \in M_1(\Sigma^2)$ is such that $Q \ll Q_1 \otimes \pi$, with $f = \frac{dQ}{dQ_1 \otimes \pi}$, then for all $x \notin N_1$ we have $Q_{2|1}(x, \cdot) \ll \pi(x, \cdot)$ with Radon-Nykodim derivative $f(x, \cdot)$, such that $\int f(x,y)\pi(x, dy) = 1$ and $N_1 \in \mathcal{B}_\Sigma$ is such that $Q_1(N_1) = 0$. The function

$$f^\pi(x,y) = \begin{cases} f(x,y) & \text{if } x \notin N_1 \\ 1 & \text{otherwise} \end{cases}$$

is then called the Borel measurable π-extension of f. Given a Markov kernel π, a π-extension of $Q_{i+1|i}(x, \cdot)$ is a Markov kernel of the form

$$Q^\pi_{i+1|i}(x, \cdot) \triangleq \begin{cases} Q_{i+1|i}(x, \cdot) & \text{if } x \notin N_i \\ \pi(x, \cdot) & \text{otherwise}, \end{cases}$$

where the Borel set N_i of zero Q_i-measure is always chosen such that if $Q_{i,i+1} \ll Q_i \otimes \pi$ then $f^\pi(x,y) = \frac{dQ^\pi_{i+1|i}(x,\cdot)}{d\pi(x,\cdot)}$ for all x. Thus, given $Q_{0,n}$, any Markov kernel π and initial distribution $p \in M_1(\Sigma)$ induce the non-homogeneous Markov measure $Q^{p,\pi} = p \otimes Q^\pi_{1|0} \otimes \cdots \otimes Q^\pi_{n|n-1}$.

The next lemma is the Markov analog of the fundamental information identity presented in [5, (2.11)] for the i.i.d. case.

Lemma 4.1 *For every $Q \in M_1(\Omega)$, every $p \in M_1(\Sigma)$, every Markov kernel π and all $n \geq 1$,*

$$H(Q_{0,n}|p \otimes_n \pi) = H(Q_{0,n}|Q^{p,\pi}) + \sum_{i=0}^{n-1} H(Q_{i,i+1}|Q_i \otimes \pi) \ . \tag{4.2}$$

Moreover, $H(Q_{0,n}|Q^{p,\pi})$ is independent of the kernel π.

Remark: The special case of Q a stationary measure with $Q_1 = p$ is derived in [22, Lemma 4.1].

Proof: Hereafter we identify Q with $Q_{0,n}$ and let $P = p \otimes_n \pi$. The value of $H(Q|Q^{p,\pi})$ is independent of π (and of the specific π-extensions we used) since

$$Q(\cup_{i=0}^{n-1}\{(x_0,\ldots,x_n) : x_i \in N_i\}) \leq \sum_{i=0}^{n-1} Q_i(N_i) = 0 \ .$$

Suppose first that $Q_{i,i+1} \ll Q_i \otimes \pi$ for $i = 0,\ldots,n-1$ and let $h_i(x,y) = \frac{dQ_{i,i+1}}{dQ_i \otimes \pi}$. By construction, $Q^{p,\pi} \ll P$ and $\frac{dQ^{p,\pi}}{dP} = \prod_{i=0}^{n-1} h_i^\pi(x_i, x_{i+1})$, where h_i^π are the π-extensions of h_i for $i = 0,\ldots,n-1$. Moreover,

$$Q(\frac{dQ^{p,\pi}}{dP} = 0) \leq \sum_{i=0}^{n-1} Q_{i,i+1}(\{(x,y) : h_i(x,y) = 0\}) = 0 \ .$$

Hence, $Q \ll Q^{p,\pi}$ iff $Q \ll P$. We thus assume that $Q \ll Q^{p,\pi} \ll P$ (otherwise (4.2) trivially holds), in which case

$$\begin{aligned} H(Q|P) &= H(Q|Q^{p,\pi}) + \int \log(\prod_{i=0}^{n-1} h_i^\pi(x_i,x_{i+1}))dQ = H(Q|Q^{p,\pi}) + \sum_{i=0}^{n-1} \int \log h_i(x_i,x_{i+1})dQ \\ &= H(Q|Q^{p,\pi}) + \sum_{i=0}^{n-1} \int_{\Sigma^2} \log h_i(x,y)dQ_{i,i+1} = H(Q|Q^{p,\pi}) + \sum_{i=0}^{n-1} H(Q_{i,i+1}|Q_i \otimes \pi) \ . \end{aligned}$$

Suppose now that for some i and some $A \in \mathcal{B}_{\Sigma^2}$, $Q_{i,i+1}(A) > 0$ while $Q_i \otimes \pi(A) = 0$. If $Q_i \not\ll P_i$ then both sides of (4.2) are ∞. Otherwise, let $A_1 = A \cap \{(x,y) : \frac{dQ_i}{dP_i}(x) > 0\}$ and observe that $Q_{i,i+1}(A_1) > 0$, while $P_i \otimes \pi(A_1) = 0$. Since $P_{i,i+1} = P_i \otimes \pi$, it thus follows that $Q_{i,i+1} \not\ll P_{i,i+1}$ implying that both sides of (4.2) are ∞. ■

The following direct corollary of Lemma 4.1 is the Markov analog of (2.5).

Corollary 4.3 *For every $Q \in M_1(\Omega)$, every $p \in M_1(\Sigma)$, every Markov kernel π and all $n \geq k \geq 1$,*

$$H(Q_{0,n}|p \otimes_n \pi) \geq \sum_{j=0}^{[n/k]-1} H(Q_{jk,(j+1)k}|Q_{jk} \otimes_k \pi) \ . \tag{4.4}$$

Proof: Let $m = [n/k]$. Since the relative entropy is non-increased by projections (see [5, (1.4)]), in particular $H(Q_{0,n}|p \otimes_n \pi) \geq H(Q_{0,mk}|p \otimes_{mk} \pi)$. For $k > 1$ fix $\nu \in M_1(\Sigma^{k-1})$ and let $\tilde{Q}$ be such that $\tilde{Q}_{0,mk+(k-1)} = \tilde{Q}_{0,k-2} \times \tilde{Q}_{k-1,mk+k-1}$, where $\tilde{Q}_{k-1,mk+k-1} = Q_{0,mk}$ and $\tilde{Q}_{0,k-2} = \nu$ (if $k = 1$ then $\tilde{Q} = Q$). Applying Lemma 4.1 (on $(\Sigma^k)^{\mathbb{Z}_+}$) for the Markov kernel $\pi_k(x, dy) = \pi(x_k, dy_1)\pi(y_1, dy_2)\cdots\pi(y_{k-1}, dy_k)$ and the initial (product) distribution $p_k = \nu \times p$, we obtain the inequality

$$H(Q_{0,mk}|p \otimes_{mk} \pi) = H(\tilde{Q}_{0,mk+(k-1)}|p_k \otimes_m \pi_k) \geq \sum_{j=0}^{m-1} H(\tilde{Q}_{jk,jk+(2k-1)}|\tilde{Q}_{jk,jk+(k-1)} \otimes \pi_k) \ .$$

Since $\pi_k(x, dy)$ is independent of $\{x_1, \ldots, x_{k-1}\}$, it follows that for all j

$$H(\tilde{Q}_{jk,jk+(2k-1)}|\tilde{Q}_{jk,jk+(k-1)}\otimes\pi_k) \geq H(\tilde{Q}_{jk+(k-1),jk+(2k-1)}|\tilde{Q}_{jk+(k-1)}\otimes\pi_k) = H(Q_{jk,(j+1)k}|Q_{jk}\otimes_k\pi) \ .$$

The proof is completed by combining the preceding inequalities. ■

Csiszár's proof of Theorem 2.1 is based on the "triangle inequality" for relative entropies (see [4, (2.14)]). The next lemma, inspired by [6, (A.1)] (in which Σ is a finite set), is the Markov analog of this inequality.

Lemma 4.5 *Let $Q, P \in M_1(\Sigma^2)$, and for $\alpha \in [0,1]$, define $R_\alpha = \alpha P + (1-\alpha)Q$. If for all $\alpha \in [0,1]$*

$$H(R_\alpha|(R_\alpha)_1 \otimes \pi) \geq H(Q|Q_1 \otimes \pi) \ , \tag{4.6}$$

then

$$H(P|P_1 \otimes \pi) \geq H(P|P_1 \otimes \overline{\pi}) + H(Q|Q_1 \otimes \pi) \ , \tag{4.7}$$

where

$$\overline{\pi}(x, \cdot) \triangleq \begin{cases} Q^{\pi}_{2|1}(x, \cdot) & \text{if } x \notin N \\ P^{\pi}_{2|1}(x, \cdot) & \text{otherwise}, \end{cases} \tag{4.8}$$

is a Markov kernel, and $N = \{x : \frac{dQ_1}{d(R_{1/2})_1}(x) = 0\} \in \mathcal{B}_\Sigma$. In particular, if $P_1 \ll Q_1$, we take $\overline{\pi} = Q^{\pi}_{2|1}$.

Proof: Without loss of generality $H(P|P_1\otimes\pi)<\infty$, and by (4.6) (for $\alpha=1$) also $H(Q|Q_1\otimes\pi)<\infty$. In particular, $Q\ll Q_1\otimes\pi$ and $P\ll P_1\otimes\pi$ with $f=\frac{dQ}{dQ_1\otimes\pi}$ and $g=\frac{dP}{dP_1\otimes\pi}$. Note that $(R_\alpha)_1\sim(R_{1/2})_1$ for all $\alpha\in(0,1)$ while both $Q_1=(R_0)_1\ll(R_{1/2})_1$ and $P_1=(R_1)_1\ll(R_{1/2})_1$. Let $m=(R_{1/2})_1\otimes\pi$ and $q_\alpha=\frac{d(R_\alpha)_1}{d(R_{1/2})_1}$ for $\alpha\in[0,1]$, i.e. $q_\alpha=\alpha q_1+(1-\alpha)q_0$ with $q_0+q_1=2$ for all $x\in\Sigma$. It is easy to check that $R_\alpha\ll m$ with $p_\alpha=\frac{dR_\alpha}{dm}=\alpha q_1 g^\pi+(1-\alpha)q_0 f^\pi$. Consequently, $H(R_\alpha|(R_\alpha)_1\otimes\pi)=\int h_\alpha dm$ where $h_\alpha=p_\alpha\log(p_\alpha/q_\alpha)$ for $\alpha\in(0,1)$. Since the mapping $\alpha\mapsto h_\alpha(x,y)$ is convex on $(0,1)$ and $h_0=\lim_{\alpha\downarrow 0}h_\alpha=p_0\log f^\pi$, $h_1=\lim_{\alpha\uparrow 1}h_\alpha=p_1\log g^\pi$, it follows that $(h_1-h_0)-(h_\alpha-h_0)/\alpha\geq 0$ is monotone non-increasing on $(0,1]$ while $H(P|P_1\otimes\pi)=\int h_1 dm<\infty$ and $H(Q|Q_1\otimes\pi)=\int h_0 dm<\infty$. Since $\int(h_\alpha-h_0)dm\geq 0$, by monotone convergence

$$\int(\lim_{\alpha\to 0}(h_\alpha-h_0)/\alpha)dm\geq 0\ .$$

Let $N=\{x:q_0(x)=0\}$ and observe that $\alpha^{-1}(h_\alpha-h_0)\to_{\alpha\to 0}p_1\log g^\pi$ for $x\in N$ and $\alpha^{-1}(h_\alpha-h_0)\to_{\alpha\to 0}q_1(g^\pi 1_{\{f^\pi>0\}}-f^\pi)+(p_1-p_0)\log f^\pi$ otherwise. Since $P_1\otimes\overline{\pi}\ll m$, with $\frac{dP_1\otimes\overline{\pi}}{dm}=q_1(1_{N^c}(x)f^\pi+1_N(x)g^\pi)$, it follows that $\int p_1 dm=1$, $\int q_1(1_{N^c}(x)f^\pi+1_N(x)g^\pi)dm=1$ and

$$H(P|P_1\otimes\overline{\pi})=-\int p_1\log(\frac{1}{p_1}\frac{dP_1\otimes\overline{\pi}}{dm})dm=-\int 1_{N^c}(x)p_1\log\frac{f^\pi}{g^\pi}dm\ .$$

Consequently,

$$\begin{aligned}
0 &\leq \int(\lim_{\alpha\to 0}(h_\alpha-h_0)/\alpha)dm\\
&= \int 1_N(x)p_1\log g^\pi dm+\int 1_{N^c}(x)(p_1-p_0)\log f^\pi dm+\int 1_{N^c}(x)(q_1 g^\pi 1_{\{f^\pi>0\}}-q_1 f^\pi)dm\\
&\leq \int p_1\log g^\pi dm-\int p_0\log f^\pi dm+\int 1_{N^c}(x)p_1\log\frac{f^\pi}{g^\pi}dm+\int 1_{N^c}(x)p_1 dm-\int 1_{N^c}(x)q_1 f^\pi dm\\
&= \int h_1 dm-\int h_0 dm+\int 1_{N^c}(x)p_1\log\frac{f^\pi}{g^\pi}dm+\int p_1 dm-\int q_1(1_N(x)g^\pi+1_{N^c}(x)f^\pi)dm
\end{aligned}$$

and the inequality (4.7) follows. ■

With a slight abuse of notation we let P_X^n denote the joint Markov law of $(X_0,X_1,\ldots,X_{n-1})$, where $X_0,X_1,\ldots,X_n,\ldots$ is a realization of the Markov chain with initial distribution p_0 and Markov transition kernel $\pi(x,dy)$ i.e. $P_X^n=p_0\otimes_{n-1}\pi$. The (modified) pair empirical measure is

$$L_{n,2}=n^{-1}(\sum_{i=0}^{n-2}\delta_{X_i,X_{(i+1)}}+\delta_{X_{n-1},X_0})\ .$$

Warning: This definition artificially introduces a transition $X_{n-1} \to X_0$ in order to assure that $L_{n,2}$ is a stationary measure.

Under suitable conditions the large deviations rate function for $L_{n,2}$ is then

$$I_2(Q|\pi) \triangleq \begin{cases} H(Q|Q_1 \otimes \pi) & \text{if } Q_1 = Q_2 \\ \infty & \text{otherwise}. \end{cases} \tag{4.9}$$

In analogy with the i.i.d. case we let $P_{L_{n,2}} \in M_1(M_1(\Sigma^2))$ denote the law of $L_{n,2}$ in $M_1(\Sigma^2)$ and whenever $P_{L_{n,2}}(\Pi') > 0$, let $P^n_{X^n|\Pi'}$ denote the law of $(X_0, X_1, \ldots, X_{n-1})$ conditioned on the event $L_{n,2} \in \Pi'$. The following assumption replaces (A–1) in the Markov setting.

Assumption (A–2) Π' *is a measurable subset of a completely convex* $\Pi \subset M_1(\Sigma^2)$, *with* $P_{L_{n,2}}(\Pi') > 0$ *and* $I_2(\Pi|\pi) = \inf_{P\in\Pi} I_2(P|\pi) < \infty$.
Here, Π' measurable means that $A_n = \{(x_0, \ldots, x_{n-1}) : n^{-1}(\sum_{i=0}^{n-2} \delta_{x_i, x_{(i+1)}} + \delta_{x_{n-1}, x_0}) \in \Pi'\} \in \mathcal{B}_{\Sigma^n}$ for all n.

The I-projection of π on Π is P^* such that $I_2(P^*|\pi) = I_2(\Pi|\pi)$ and $\alpha Q + (1-\alpha)P^* \in \Pi$ for every $Q \in \Pi$. Note that as defined here, the I-projection of π on Π might not exist, and even when $P^* \in \Pi$ exists, it might be non-unique (see [6, Example 2]). Given P^*, let $f = \frac{dP^*}{dP_1^* \otimes \pi}$, and construct its π-extension $f^\pi(x,y)$. The Markov kernel $\pi^*(x, dy) \triangleq f^\pi(x,y)\pi(x,dy)$ induces the Markov measure $p_0 \otimes_{n-1} \pi^*$ on $M_1(\Sigma^n)$ to be denoted $(P^*)^n$. Note that $(P^*)^n \ll P^n_X$ with

$$\frac{d(P^*)^n}{dP^n_X} = \prod_{i=0}^{n-2} f^\pi(x_i, x_{i+1}) \, .$$

In what follows, $Q_{n-1,0}$ denotes the marginal of a measure $Q \in M_1(\Sigma^n)$ on its last and first coordinates, that is, for any Borel set $B \subset \Sigma^2$, $Q_{n-1,0}(B) = Q(\{(x_0, \ldots, x_{n-1}) : (x_{n-1}, x_0) \in B\})$.

Proposition 4.10 *Assume (A–2). Let* $Q = P^n_{X^n|\Pi'}$. *Suppose that* $H(Q_{n-1,0}|Q_{n-1} \otimes \pi) < \infty$, *and that there exists an* I*-projection* P^* *of* π *on* Π *such that* $n^{-1}\sum_{i=0}^{n-1} Q_i \ll P_1^*$. *Then,*

$$\frac{1}{n} H(P^n_{X^n|\Pi'}|(P^*)^n) \leq -\frac{1}{n}\log P_{L_{n,2}}(\Pi') - I_2(\Pi|\pi) + \frac{1}{n}\int \log f^\pi \, dQ_{n-1,0} \, . \tag{4.11}$$

Proof: Since $Q(\Gamma) = P^n_X(\Gamma \cap A_n)/P^n_X(A_n)$ for all $\Gamma \in \mathcal{B}_{\Sigma^n}$, it follows that $Q \ll P^n_X$ with

$$H(Q|P^n_X) = -\log P^n_X(A_n) = -\log P_{L_{n,2}}(\Pi') < \infty \, .$$

Applying (4.2) for $p = p_0$ and the Markov kernels π^* we obtain the identity

$$H(Q|(P^*)^n) = H(Q|Q^{p_0,\pi^*}) + \sum_{i=0}^{n-2} H(Q_{i,i+1}|Q_i \otimes \pi^*) .$$

Applying again (4.2), now with the Markov kernel π, we also have

$$-\log P_{L_{n,2}}(\Pi') = H(Q|Q^{p_0,\pi}) + \sum_{i=0}^{n-2} H(Q_{i,i+1}|Q_i \otimes \pi)$$

(in particular, the right side is finite), and since $H(Q|Q^{p_0,\pi}) = H(Q|Q^{p_0,\pi^*})$, it follows that

$$H(P^n_{X^n|\Pi'}|(P^*)^n) + \log P_{L_{n,2}}(\Pi') = \sum_{i=0}^{n-2}[H(Q_{i,i+1}|Q_i \otimes \pi^*) - H(Q_{i,i+1}|Q_i \otimes \pi)] .$$

Since $H(Q_{i,i+1}|Q_i \otimes \pi)$ is finite and $f^\pi = \frac{dQ_i \otimes \pi^*}{dQ_i \otimes \pi}$, it follows that

$$H(Q_{i,i+1}|Q_i \otimes \pi^*) = H(Q_{i,i+1}|Q_i \otimes \pi) + \int \log(\frac{1}{f^\pi}) dQ_{i,i+1}$$

with the integral well defined, but possibly infinite (see [4, (2.6)]). We show below that

$$\int \log f^\pi \, d\tilde{P} \geq \int \log f^\pi \, dP^* = I_2(\Pi|\pi) \geq 0 \tag{4.12}$$

where $\tilde{P} = n^{-1} \sum_{i=0}^{n-1} Q_{i,(i+1)\bmod n}$ and the left-side is well defined, but possibly infinite. Thus,

$$\begin{aligned} H(P^n_{X^n|\Pi'}|(P^*)^n) + \log P_{L_{n,2}}(\Pi') + nI_2(\Pi|\pi) &\leq n \int \log f^\pi \, d\tilde{P} + \sum_{i=0}^{n-2} \int \log(\frac{1}{f^\pi}) \, dQ_{i,i+1} \\ &= \int \log f^\pi \, dQ_{n-1,0} . \end{aligned}$$

It remains to prove (4.12). To this end, recall that for every $\Gamma \in \mathcal{B}^2_\Sigma$,

$$\tilde{P}(\Gamma) = n^{-1} \sum_{i=0}^{n-1} P((X_i, X_{(i+1)\bmod n}) \in \Gamma | L_{n,2} \in \Pi') = E[L_{n,2}(\Gamma)|L_{n,2} \in \Pi'] = E[L_{n,2}|L_{n,2} \in \Pi'](\Gamma) ,$$

implying that $\tilde{P} \in \Pi$ with $\tilde{P}_2 = \tilde{P}_1$. Note that for $\alpha \in (0,1]$, both $\tilde{P}_\alpha = \alpha\tilde{P} + (1-\alpha)P^* \in \Pi$ and $(\tilde{P}_\alpha)_1 = (\tilde{P}_\alpha)_2$. Since P^* is the I-projection of π on Π, it follows that $H(P^*|P^*_1 \otimes \pi) \leq H(\tilde{P}_\alpha|(\tilde{P}_\alpha)_1 \otimes \pi)$. Therefore, by Lemma 4.5

$$H(\tilde{P}|\tilde{P}_1 \otimes \pi) \geq H(\tilde{P}|\tilde{P}_1 \otimes \overline{\pi}) + H(P^*|P^*_1 \otimes \pi) ,$$

where $\overline{\pi}$ is the Markov kernel of (4.8). Since $\tilde{P}_1 \ll P_1^*$, $\overline{\pi} = \pi^*$ and in particular, $f^\pi = \frac{d\tilde{P}_1 \otimes \overline{\pi}}{d\tilde{P}_1 \otimes \pi}$. Since $H(Q_{n-1,0}|Q_{n-1} \otimes \pi) < \infty$, it follows by the convexity of $H(\cdot|(\cdot)_1 \otimes \pi)$ that also $H(\tilde{P}|\tilde{P}_1 \otimes \pi) < \infty$. Consequently, by [4, (2.6)]

$$H(\tilde{P}|\tilde{P}_1 \otimes \pi) - H(\tilde{P}|\tilde{P}_1 \otimes \overline{\pi}) = \int \log f^\pi \, d\tilde{P} \, ,$$

yielding (4.12). ■

For any $0 \leq \ell < \ell' \leq n$, let $P^n_{X_\ell^{\ell'}|\Pi'}$ denote the law of $(X_\ell, X_{\ell+1}, \ldots, X_{\ell'-1})$ conditioned on the event $L_{n,2} \in \Pi'$. A direct consequence of Proposition 4.10 and Corollary 4.3 is the following

Corollary 4.13 *Assume the conditions of Proposition 4.10. Further assume that for some $k(n)$,*

$$-\frac{k(n)}{n} \log P_{L_{n,2}}(\Pi') - k(n) I_2(\Pi|\pi) + \frac{k(n)}{n} \int \log f^\pi \, dQ_{n-1,0} \to_{n\to\infty} 0 \, . \tag{4.14}$$

Let $J \in \{0, 1, \ldots, [n/k(n)] - 1\}$ be uniformly distributed. Then

$$H(P^n_{X_{Jk(n)}^{(J+1)k(n)}|\Pi'} | P^n_{X_{Jk(n)}^{Jk(n)+1}|\Pi'} \otimes_{k(n)} \pi^*) \to_{n\to\infty} 0 \text{ in } L_1 \, . \tag{4.15}$$

That is, a block of length $k(n)$ whose starting point is distributed uniformly over the sequence behaves, under the conditioning, like a Markov chain with transition kernel π^*.

We conclude by describing one concrete example, a Markov counterpart of Corollary 2.7. Assume there exists a dominating measure $m \in M_1(\Sigma)$ such that, for some $b > 1$,

$$bm(\cdot) \geq \pi(x, \cdot) \geq b^{-1} m(\cdot), \quad H(p_0|m) < \infty \, . \tag{4.16}$$

Let μ_π be the invariant measure for the kernel π and $\psi : \Sigma \to \mathbb{R}^\ell$ a Borel measurable map with $K = \sup_y |\psi(y)| < \infty$.

Corollary 4.17 *Suppose that the law of $\psi(X_1)$ conditioned on $X_0 = x$, is either strongly non-lattice for m-a.e. x, or is lattice with the same span for m-a.e. x. Let $C \subset \mathbb{R}^\ell$ be a convex set such that $\int \psi(y) \mu_\pi(dy) \notin \overline{C}$ and C^o intersects the interior of the convex hull of the support of $m_\psi = m \circ \psi^{-1}$. Then, (4.15) holds true for $\Pi' = \{\nu : \int \psi d\nu_2 \in C\}$ and $n^{-1} k(n) \log n \to 0$.*

Remark: Note that $L_{n,2} \in \Pi'$ is equivalent to the statement that $n^{-1} \sum_{i=0}^{n-1} \psi(X_i) \in C$.

Proof: Π' is a measurable subset of the completely convex $\Pi = \{\nu : \int \psi d\nu_2 \in \overline{C}\} \in M_1(\Sigma^2)$ and

$$P_{L_{n,2}}(\Pi') = P(\frac{1}{n}\sum_{i=0}^{n-1}\psi(X_i) \in C) .$$

Our assumption (4.16) translates into [16, condition (3.1)] and since ψ is bounded [16, Theorem 5.3] applies, yielding for some $\eta < \infty$ and $0 < c_1 < c_2 < \infty$,

$$c_1 n^{-\ell/2} e^{-n\eta} \le P_{L_{n,2}}(\Pi') \le c_2 e^{-n\eta} . \tag{4.18}$$

For every $\alpha \in \mathbb{R}^\ell$, by [14, Theorem III.10.1], the kernel $\pi_\alpha(x, dy) = \pi(x, dy)e^{\langle \alpha, \psi(y)\rangle}$ has a maximal simple positive eigenvalue $e^{\Lambda(\alpha)}$, with associated (right) eigenfunction r_α and (left) eigenmeasure σ_α such that r_α and $\frac{d\sigma_\alpha}{dm}$ are bounded and uniformly positive and $\int r_\alpha d\sigma_\alpha = 1$. The probability measure

$$P_\alpha(dx, dy) = e^{-\Lambda(\alpha)}\sigma_\alpha(dx)\pi_\alpha(x, dy)r_\alpha(y) ,$$

has marginals $(P_\alpha)_1 = (P_\alpha)_2 = r_\alpha\sigma_\alpha \gg m$, with

$$\log \frac{dP_\alpha}{d(P_\alpha)_1 \otimes \pi}(x, y) = \langle \alpha, \psi(y)\rangle - \Lambda(\alpha) + \log r_\alpha(y) - \log r_\alpha(x)$$

a bounded function. By [16, Corollary 4.1], $\nabla\Lambda(\alpha) = \int \psi d(r_\alpha\sigma_\alpha)$ and hence

$$I_2(P_\alpha|\pi) = \langle \alpha, \int \psi d(P_\alpha)_2\rangle - \Lambda(\alpha) = \langle \alpha, \nabla\Lambda(\alpha)\rangle - \Lambda(\alpha) . \tag{4.19}$$

By [16, Theorem 5.2 and (4.27)], $\eta = I_2(P_{\alpha_o}|\pi) < \infty$ for some α_o such that $P^* = P_{\alpha_o} \in \Pi$. In particular, assumption (A–2) holds.

Applying [7, Theorem 10] to the τ-open set $\widehat{\Pi} = \{\nu : \int \psi d\nu_2 \in C^o\} \subset \Pi'$, we obtain that

$$-I_2(P^*|\pi) = -\eta \ge \liminf_{n\to\infty} \frac{1}{n} \log P_{L_{n,2}}(\widehat{\Pi}) \ge -\inf_{Q\in\widehat{\Pi}} I_2(Q|\pi) . \tag{4.20}$$

Our assumptions imply (again, by [16, Theorem 5.2 and (4.27)] applied to a closed, convex subset of C^o) that there exists α' such that $Q = P_{\alpha'} \in \widehat{\Pi}$ with $I_2(Q|\pi) < \infty$ (see (4.19)). Since $Q_\alpha = \alpha Q + (1-\alpha)P \in \widehat{\Pi}$ for any $P \in \Pi$ and all $\alpha \in (0, 1]$, we consequently obtain by (4.20) and the convexity of $I_2(\cdot|\pi)$ that

$$I_2(P^*|\pi) \le I_2(Q_\alpha|\pi) \le \alpha I_2(Q|\pi) + (1-\alpha)I_2(P|\pi) \to_{\alpha\to 0} I_2(P|\pi) .$$

Therefore, P^* is the I-projection of π on Π, and condition (4.14) holds (see (4.18)).

Turning to check that all assumptions of Proposition 4.10 hold, note first that (4.16) obviously implies that $Q_i \ll m$, hence also $\frac{1}{n}\sum_{i=0}^{n-1} Q_i \ll m \ll P_1^*$. Let $C^{-\delta} = \{x : \inf_{y\notin C} |x-y| > \delta\}$. Clearly, for any Borel set $B \subset \Sigma^2$,

$$Q_{n-1,0}(B) P_{L_{n,2}}(\Pi') \leq P((X_{n-1}, X_0) \in B) \leq b^{n-1} m \times p_0(B) ,$$

while

$$\begin{aligned} Q_{n-1} \otimes \pi(B) P_{L_{n,2}}(\Pi') &\geq P((X_{n-1}, X_n) \in B, \frac{1}{n}\sum_{i=1}^{n-2} \psi(X_i) \in C^{-2K/n}) \\ &\geq b^{-n} m \times m(B) m_\psi^{n-2}(\{(v_1, \ldots, v_{n-2}) : \frac{1}{n}\sum_{i=1}^{n-2} v_i \in C^{-2K/n}\}) . \end{aligned}$$

It follows from Cramèr's theorem and the support condition on m_ψ that $m_\psi^{n-2}(\{(v_1, \ldots, v_{n-2}) : \frac{1}{n}\sum_{i=1}^{n-2} v_i \in C^{-2K/n}\}) > 0$ for all n large enough. Hence, $\frac{dQ_{n-1,0}}{dQ_{n-1}\otimes\pi}(x,y) \leq c_n \frac{dp_0}{dm}(y)$ for some constant c_n which is finite for all n large enough, leading by (4.16) to $H(Q_{n-1,0}|Q_{n-1}\otimes\pi) < \infty$. We conclude by applying Corollary 4.13. ■

Acknowledgment We thank P. Diaconis for several discussions concerning [9]. We also thank P. Bickel and A. Mandelbaum for guiding us through the U-statistics literature.

References

[1] R. N. Bhattacharya and R. Ranga Rao *Normal approximation and asymptotic expansions.* Wiley, New York, 1976.

[2] E. Bolthausen. Laplace approximations for sums of independent random vectors, part II. *Probab. Th. Rel. Fields*, 76:167–206, 1987.

[3] E. Bolthausen. Maximum entropy principles for Markov processes. *Stochastic processes and applications in mathematical physics. Proceeding of the Bielefeld Conference, 1985.* Mathematics and its applications, Vol. 61. Kluwer, 1990.

[4] I. Csiszár. *I*–divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

[5] I. Csiszár. Sanov property, generalized I–projection and a conditional limit theorem. *Ann. Probab.*, 12:768–793, 1984.

[6] I. Csiszár, T. M. Cover, and B. S. Choi. Conditional limit theorems under Markov conditioning. *IEEE Trans. Inf. Theory*, 33:788–801, 1987.

[7] A. De Acosta. Large deviations for empirical measures of Markov chains. *J. Theor. Probab.* 3:395–431, 1990.

[8] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[9] P. Diaconis and D. Freedman. Finite exchangeable sequences. *Ann. Probab.*, 4:745–764, 1980.

[10] P. Diaconis and D. Freedman. Conditional limit theorems for exponential families and finite versions of de Finetti's theorem. *J. Theoretical Prob.*, 1:381–410, 1988.

[11] N. Dunford and J. T. Schwartz. *Linear Operators, Part I.* Interscience Publishers Inc., New York, 1958.

[12] I. H. Dinwoodie. Measures dominantes et théorème de Sanov. *Ann. Inst. Henry Poincaré*, 28:365–373, 1992.

[13] R. M. Dudley, E. Giné and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *J. of Theoretical Probability* 4:485–510, 1991.

[14] T. E. Harris. *The Theory of Branching Processes* Springer-Verlag, 1963, Berlin.

[15] M. Iltis. Asymptotics of Large Deviations in $\mathbb{R}^d$. Preprint, 1993.

[16] I. Iscoe, P. Ney and E. Nummelin. Large deviations of uniformly recurrent Markov additive processes. *Adv. in Appl. Math.* 6:373–412, 1985.

[17] V. S. Korolyuk and Yu. V. Borovskikh. Convergence rate for degenerate Von Mises functionals. *Theory Probab. Appl.*, 33:125–135, 1988

[18] O.E. Lanford. Entropy and equilibrium states in classical statistical mechanics. Lecture notes in Physics Vol. 20, 1–113, 1973. Springer, New York.

[19] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer-Verlag, 1991, New York.

[20] P. Ney. Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$. *Ann. Probab.*, 11:158–167, 1983.

[21] D. Pollard *Convergence of stochastic processes*. Springer-Verlag, 1984, New York.

[22] C. Schroeder. I-projection and conditional limit theorems for discrete parameter Markov processes. *Ann. Probab.*, 21:721–758, 1993.

[23] D. W. Stroock and O. Zeitouni. Microcanonical distributions, Gibbs states, and the equivalence of ensembles. In R. Durrett and H. Kesten, editors, *Festchrift in honor of F. Spitzer*, pages 399–424. Birkhäuser, Basel, Switzerland, 1991.

[24] B. von Bahr and C.G. Esseen. Inequalities for the r-th absolute moment for a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statistics*, 36:299–303, 1965.