The Annals of Probability
2000, Vol. 28, No. 1, 416–461

CONCENTRATION OF MEASURE INEQUALITIES FOR MARKOV CHAINS AND Φ-MIXING PROCESSES

BY PAUL-MARIE SAMSON

University of Toulouse

We prove concentration inequalities for some classes of Markov chains and Φ-mixing processes, with constants independent of the size of the sample, that extend the inequalities for product measures of Talagrand. The method is based on information inequalities put forward by Marton in case of contracting Markov chains. Using a simple duality argument on entropy, our results also include the family of logarithmic Sobolev inequalities for convex functions. Applications to bounds on supremum of dependent empirical processes complete this work.

1. Introduction. In a recent series of striking papers (see [15], [16], [17]), Talagrand deeply analyzed the concentration of measure phenomenon in product space, with applications to various areas of probability theory. A first result at the origin of his investigation concerns deviation inequalities for product measures $P = \mu_1 \otimes \cdots \otimes \mu_n$ on $[0, 1]^n$. Namely, for every convex function f on $[0, 1]^n$, with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leq 1$, and for every $t \geq 0$,

$$P(|f - M| \geq t) \leq 4 \exp\left(-\frac{t^2}{4}\right), \tag{1.1}$$

where M is a median of f for P. This Gaussian-type bound may be considered as an important generalization of the classical inequalities for sums of independent random variables. The deviation inequality (1.1) is a consequence of a concentration inequality on sets which takes the following form. To measure the "distance" of a point $x \in \mathbb{R}^n$ to a set A, consider the functional (see "convex hull," [15], Chapter 4),

$$f_{\mathrm{conv}}(A, x) = \sup_{\alpha} \inf_{y \in A} \left(\sum_{i=1}^{n} \alpha_i \mathbf{1}_{x_i \neq y_i} \right),$$

where the supremum is over all vectors $\alpha = (\alpha_i)_{1 \leq i \leq n}$, $\alpha_i \geq 0$, $\sum_{i=1}^n \alpha_i^2 = 1$. If we let $A_t^{\mathrm{conv}} = \{x \in \mathbb{R}^n, f_{\mathrm{conv}}(A, x) \leq t\}$, Talagrand shows that for every $t \geq \sqrt{2 \log(1/P(A))}$,

$$P\big(A_t^{\mathrm{conv}}\big) \geq 1 - \exp\left[-\frac{1}{2} \left(t - \sqrt{2 \log \frac{1}{P(A)}} \right)^2 \right]. \tag{1.2}$$

Besides the convex hull approximation, Talagrand considers two other approximations on product spaces for which he proves similar concentration proper-

Received April 1998.

ties. One of the main features of these inequalities is that they are independent of the dimension of the product space, that is, of the size of the sample. We will be mainly concerned with extensions of the convex hull approximation in this work.

Recently, an alternate, simpler, approach to some of Talagrand's inequalities was suggested by Ledoux [7] on the basis of log-Sobolev inequalities. Introduce, for every function g on $\mathbb{R}^n$, the entropy functional,

$$\mathrm{Ent}_P\left(g^2\right) = \int g^2 \log\, g^2\, dP - \int g^2\, dP \,\log \int g^2\, dP.$$

Then, it can easily be shown that, for every product measure P on $[0, 1]^n$ and for every separately convex function f,

$$\mathrm{Ent}_P\left(e^f\right) \le \tfrac{1}{2} \int |\nabla f|^2\, e^f\, dP,$$

where ∇f denotes the usual gradient of f on $\mathbb{R}^n$ and $|\nabla f|$ its Euclidean length. This inequality easily implies deviation inequalities of the type of (1.1). Indeed, the preceding log-Sobolev inequality may be turned into a differential inequality on the Laplace transform of convex Lipschitz functions, which then yields tail estimates by Chebyshev's inequality. This type of argument may be pushed further to recover most of Talagrand's deviation inequalities for functions [7]. It however does not seem to succeed for deviations under the median (or for concave functions).

A third approach to concentration for product measures was developed by Marton [8] using inequalities from information theory. This method, which lies at the level of measures rather than sets or functions and also uses entropic inequalities, allows her to recover Talagrand's convex hull concentration (1.2). Dembo [3] further developed this line of reasoning to reach the other types of approximations in product spaces introduced by Talagrand (see also [4]). Besides describing a new method of proof, Marton's approach is moreover well suited to extensions to some dependent situations such as contracting Markov chains.

The main purpose of this work is to extend Marton's information theoretic approach to larger classes of dependent sequences such as Doeblin recurrent Markov chains [13] and Φ-mixing processes [5]. Φ-mixing coefficients have been recently introduced by Marton to control dependence and prove concentration inequalitites with the Hamming distance for dependent sequences (see [10]). Let, for example, $(X_i)_{i\in\mathbb{Z}}$ be a Markov chain or a Φ-mixing process. Denote by P the law on $\mathbb{R}^n$ of a sample X of size n taken from $(X_i)_{i\in\mathbb{Z}}$. We will introduce a matrix Γ of dimension n, with coefficients that will measure the dependence between the random variables $(X_1, \dots, X_n)$ of the sample X. In the interesting cases, the operator norm $\|\Gamma\|$ of the matrix Γ will be bounded independently of the size of the sample. This condition is satisfied for contracting Markov chains (see [8]), but also for more useful processes. Examples include uniformly ergodic Markov chains (see [13]) satisfying the so-called Doeblin condition (see Proposition 1). Other examples are the Φ-mixing processes for which the sequence of Φ-mixing coefficients is summable, for example, Φ-mixing processes with a geometric decay of their Φ-mixing coefficients (see [5]). All these examples are described at the beginning of Section 2.

Let now P denote the law of the sample X on $\mathbb{R}^n$. For every probability measures Q and R on $\mathbb{R}^n$, let $\mathscr{M}(Q, R)$ denote the set of all probability measures on $\mathbb{R}^n \otimes \mathbb{R}^n$ with marginals Q and R. Define

$$d_2(Q, R) = \inf_{\Pi \in \mathscr{M}(Q, R)} \sup_{\alpha} \iint \sum_{i=1}^{n} \alpha_i(y) \mathbf{1}_{x_i \neq y_i} \, d\Pi(x, y),$$

where the $\sup_\alpha$ is over all vectors of positive functions $\alpha = (\alpha_1, \dots, \alpha_n)$, with

$$\int \sum_{i=1}^{n} \alpha_i^2(y) \, dR(y) \leq 1.$$

As a main result, we show in Theorem 1 below that, for every probability measure Q on $\mathbb{R}^n$ with Radon–Nikodym derivative dQ/dP with respect to the measure P,

$$d_2(Q, P) \leq \|\Gamma\| \sqrt{2 \operatorname{Ent}_P\left(\frac{dQ}{dP}\right)}.$$

Furthermore,

$$d_2(P, Q) \leq \|\Gamma\| \sqrt{2 \operatorname{Ent}_P\left(\frac{dQ}{dP}\right)}.$$

Such Pinsker type inequalities have already been investigated by Marton for contracting Markov chains [8], and then by Dembo in the independent case [3]. Recently, Marton also obtained related bounds with a parameter readily comparable to $\|\Gamma\|$ [11]. Following these works, we could easily derive concentration in the form of (1.2) [and thus (1.1)] from these information inequalities. We however take a somewhat different route related to exponential integrability and log-Sobolev inequalities. Actually, to get concentration inequalities around the mean with the best constant (see Corollary 3), we adapt a duality argument by Bobkov and Götze [2] dealing with the equivalence between exponential inequalities on the Laplace transform and information inequalities. Let P denote the law of a sample $(X_1, \dots, X_n)$ of bounded random variables $0 \leq X_i \leq 1$. We will obtain deviation inequalities which include Berstein-type inequalities. Namely, for every Lipschitz convex function f on $[0, 1]^n$, with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leq 1$ and every $t \geq 0$,

$$P(|f - \mathrm{E}_P(f)| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\|\Gamma\|^2}\right). \tag{1.3}$$

Following this approach, we get in the same way some new log-Sobolev inequalities (see Corollary 1). From these inequalities, we could also obtain deviation inequalities such as (1.3) by the log-Sobolev method suggested by Ledoux. Nevertheless, we get a worse constant $8\|\Gamma\|^2$ instead of $2\|\Gamma\|^2$ in (1.3).

Let us note that the constant $2\|\Gamma\|^2$ is optimal as can be seen from the central limit theorem in the independent case ($\|\Gamma\| = 1$).

In Section 3, we present some applications of Theorem 1 to empirical processes, in particular to tail estimates for the supremum of empirical processes. Let S be a measurable space and let $X = (X_1, \dots, X_n)$ be a sample of random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ taking values in S. For example, X could be taken out of a sequence $(X_i)_{i\in\mathbb{Z}}$ which is a uniformly ergodic Markov chain or a Φ-mixing process. Let $\mathscr{F}$ be a countable family of bounded measurable functions g on S, $|g| \le C$. Let Z denote the random variable

$$Z = \sup_{g\in\mathscr{F}} \left| \sum_{i=1}^{n} g(X_i) \right|.$$

In the independent case, Talagrand proved sharp bounds on the tail of Z around its mean that extend the classical real-valued setting (see Theorem 1.4, [17]). More precisely, he showed that for every $t \ge 0$,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \ge t) \le K \exp\left(-\frac{1}{K}\frac{t}{C} \log\left(1 + \frac{Ct}{\mathbb{E}(\Sigma^2)} \right) \right), \tag{1.4}$$

where K is a numerical constant and

$$\Sigma^2 = \sup_{g\in\mathscr{F}} \sum_{i=1}^{n} g^2(X_i).$$

If one is only interested in bounds on $\mathbb{P}(Z \ge t + \mathbb{E}(Z))$ above the mean, the log-Sobolev method of [1] provides an efficient way to prove inequalities such as (1.4) with a simplicity that contrasts with the argument of [17]. Sharp constants in Ledoux's method have been recently obtained by Massart [12]. For us, it will be more convenient to deduce deviation inequalities for empirical processes from the information inequalities of Theorem 1. The method we will use is still linked to the equivalence between exponential integrability and information inequalities. However, we will only prove the Gaussian bound for small t's in (1.4), and we do not succeed in proving the Poissonian bound for large t's in this context of dependence. Our results are of some interest when the functions g of $\mathscr{F}$ are nonnegative (see Theorem 2). Nethertheless, in the case of arbitrary bounded functions, we could expect some improvement of the deviation inequalities of Theorem 3 (this point is developed in the Section 3).

2. Information inequalities for processes and Log-Sobolev inequalities. In this section, we present the central result of this work. On some probability space $(\Omega, \mathscr{A}, \mathbb{P})$, consider a sample $X = (X_1, \dots, X_n)$ of real-valued random variables.

As described in the introduction, the case of independent X_i's, or of a product measure P, has been extensively investigated in recent years. We are interested here in a sample X of random variables which are not necessarily independent. For example, the random variables $X_1, \dots, X_n$ of the sample X are taken out of a sequence $(X_i)_{i\in\mathbb{Z}}$ which is a Markov chain.

To measure the dependence between the random variables $X_1, \dots, X_n$, we define a triangular matrix $\Gamma = (\gamma_i^j)_{1\le i,\, j\le n}$. For $i \ge j$,

$$\gamma_i^j = \begin{cases} 0, & \text{if } i > j, \\ 1, & \text{if } i = j. \end{cases}$$

For $1 \le i < j \le n$, let X_i^j represent the vector $(X_i, \dots, X_j)$, and let

$$\mathscr{L}(X_j^n | X_1^{i-1} = y_1^{i-1},\ X_i = x_i)$$

denote the law of X_j^n conditionally to $X_1^{i-1} = y_1^{i-1}$ and $X_i = x_i$. For every $1 \le i < j \le n$ and for $x_i, y_1, \dots, y_i$ in $\mathbb{R}$, let

$$\begin{aligned} a_j(y_1^{i-1}, x_i, y_i) = \big\| \mathscr{L}(X_j^n | X_1^{i-1} &= y_1^{i-1},\ X_i = x_i) \\ &- \mathscr{L}(X_j^n | X_1^{i-1} = y_1^{i-1},\ X_i = y_i) \big\|_{\mathrm{TV}}, \end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation of a signed measure. Set then

$$(\gamma_i^j)^2 = \sup_{(x_i, y_i)\in\mathbb{R}^2} \sup_{y_1^{i-1}\in\mathbb{R}^{i-1}} a_j(y_1^{i-1}, x_i, y_i). \tag{2.1}$$

To avoid the strong condition imposed by the supremum in the definition of γ_i^j, we consider another possible definition for the coefficients of the triangular matrix Γ. For every $1 \le i < j \le n$, let

$$\tilde{a}_j(y_1^i) = \big\| \mathscr{L}(X_j^n | X_1^i = y_1^i) - \mathscr{L}(X_j^n) \big\|_{\mathrm{TV}}$$

and

$$(\tilde{\gamma}_i^j)^2 = 2 \operatorname*{ess\,sup}_{y_1^i\in\mathbb{R}^i,\, \mathscr{L}(x_1^i)} \tilde{a}_j(y_1^i), \tag{2.2}$$

where $\operatorname{ess\,sup}_{y_1^i\in\mathbb{R}^i,\, \mathscr{L}(x_1^i)}$ is the essential supremum with respect to the measure $\mathscr{L}(X_1^i)$. By definition, for every measurable function a on a probability space $(E, \mathscr{E}, \mu)$,

$$\operatorname*{ess\,sup}_{y\in E,\, \mu} a(y) = \inf\{\alpha \in \mathbb{R}^+ \cup \{\infty\},\ \mu(a(y) > \alpha) = 0\}.$$

Now, consider $\|\Gamma\|$, the usual operator norm of the matrix Γ with respect to the Euclidean topology. $\|\Gamma\|$ appears in all the results we present in our paper. Roughly speaking, it measures the "L^2-dependence" of the random variables $X_1, \dots, X_n$.

Our main emphasis will be to describe cases for which $\|\Gamma\|$ may be bounded independently of n, the size of the sample (as is of course the case when the X_i's are independent, for which $\Gamma = \mathrm{Id}$, and $\|\Gamma\| = 1$). Let us describe a few examples of interest.

A first class of examples concerns Markov chains. Assume $X_1, \dots, X_n$ is a Markov chain. By the Markov property, the coefficients γ_i^j or $\tilde{\gamma}_i^j$ take a simpler form. Namely, for $1 \le i < j \le n$,

$$(\gamma_i^j)^2 = \sup_{(x_i, y_i) \in \mathbb{R}^2} \|\mathscr{L}(X_j | X_i = x_i) - \mathscr{L}(X_j | X_i = y_i)\|_{\mathrm{TV}} \tag{2.3}$$

and

$$(\tilde{\gamma}_i^j)^2 = 2 \operatorname*{ess\,sup}_{y_i \in \mathbb{R}, \mathscr{L}(X_i)} \|\mathscr{L}(X_j | X_i = y_i) - \mathscr{L}(X_j)\|_{\mathrm{TV}}.$$

There are many examples of Markov chains for which $\|\Gamma\|$ is bounded independently of the dimension n. Let us briefly present two of them.

We first mention the Doeblin recurrent Markov chains presented, for example, in [5] (see page 88). Let $X_1, \dots, X_n$ be a homogeneous Markov chain with transition kernel $K(\cdot\,|\,\cdot)$ (for every $2 \le i \le n$, $\mathscr{L}(X_i | X_{i-1} = x_{i-1}) = K(\cdot\,|x_{i-1})$). Let μ be some nonnegative measure with nonzero mass μ_0. The next statement is due to Ueno and Davidov (see [5], page 88).

PROPOSITION 1. *If there exists some integer r such that for all x_1 in $\mathbb{R}$ and all measurable sets A,*

$$K^r(A|x_1) \le \mu(A),$$

then, for every integer k and for every x_1, y_1 in $\mathbb{R}$,

$$\|K^k(\cdot\,|x_1) - K^k(\cdot\,|y_1)\|_{\mathrm{TV}} \le 2\rho^{k/r}, \tag{2.4}$$

where $\rho = 1 - \mu_0$.

Markov chains for which the k-step transition kernels K^k satisfy (2.4) are called uniformly ergodic in [13] (see Chapter 16). In this book, there are several conditions equivalent to (2.4), in particular the so-called Doeblin condition (cf. [13], Theorem 16.0.2). The above proposition simply follows from Theorem 16.2.4 in [13]. In [5], Doukhan gives the analogue of Proposition 1 for nonhomogeneous Markov chains (cf. page 88).

If the Markov chain satisfies Proposition 1, it may be shown that

$$\|\Gamma\| \le \frac{\sqrt{2}}{1 - \rho^{1/2r}}. \tag{2.5}$$

Indeed, according to the definition (2.3) of γ_i^j, for $1 \le i < j \le n$,

$$(\gamma_i^j)^2 = \sup_{(x_i, y_i) \in \mathbb{R}^2} \|K^{j-i}(\cdot\,|x_i) - K^{j-i}(\cdot\,|y_i)\|_{\mathrm{TV}}$$

Therefore, by (2.4), for $1 \le i < j \le n$,

$$\gamma_i^j \le \sqrt{2}(\rho^{1/2r})^{j-i}. \tag{2.6}$$

Consequently,

$$\|\Gamma\| \le \sqrt{2}\left\|\mathrm{Id} + \sum_{k=1}^{n-1}(\rho^{1/2r})^k N_k\right\|,$$

where $N_k = \left(n_{ij}^{(k)}\right)_{1\le i,\, j\le n}$ represents the nilpotent matrix of order k defined by

$$n_{ij}^{(k)} = \begin{cases} 1, & \text{if } j-i=k,\\ 0, & \text{otherwise.}\end{cases}$$

Since for each $1 \le k \le n$, $\|N_k\| \le 1$, it follows from the triangular inequality that

$$\|\Gamma\| \le \sqrt{2}\sum_{k=1}^{n-1}(\rho^{1/2r})^k.$$

Finally, the geometric sum on the right-hand side is bounded independently of n, since $\rho < 1$. We thus obtain (2.5).

A second class of Markov chains is called "contracting" Markov chains in [8]. These Markov chains are not necessary homogeneous. As we already mentioned in the introduction, Marton obtains a concentration inequality for those Markov chains. This result is equivalent to our deviation inequality (2.20) in Corollary 4 applied to this particular case of Markov chains. Let K_i denote the transition kernel at the step i. In other words, $K_i(\cdot\,|x_{i-1})$ denotes the law of X_i given $X_{i-1} = x_{i-1}$. The chain will be called contracting if for every $i = 1,\dots,n$,

$$\alpha_i = \sup_{(x_{i-1},\, y_{i-1})\in\mathbb{R}^2} \left\|K_i(\cdot\,|y_{i-1}) - K_i(\cdot\,|x_{i-1})\right\|_{\mathrm{TV}} < 1. \tag{2.7}$$

In this case, $\|\Gamma\|$ may also be bounded independently of the dimension n as

$$\|\Gamma\| \le \frac{1}{1-\alpha^{1/2}}, \tag{2.8}$$

where

$$\alpha = \max_{1\le i\le n}\alpha_i.$$

To prove inequality (2.8), we first show that for every $1 \le i < j \le n$,

$$\left(\gamma_i^j\right)^2 \le \prod_{l=i+1}^{j}\alpha_l \le \alpha^{j-i}. \tag{2.9}$$

Then, replacing $\beta^{1/2r}$ by $\alpha^{1/2}$ in (2.6), the conclusion follows as in the previous example. The proof of (2.9) below is of particular interest since we will mention there a recurring argument throughout this paper. For every $2 \le i \le n$, define

$$b_i(x_{i-1}, y_{i-1}) = \left\|K_i(\cdot\,|y_{i-1}) - K_i(\cdot\,|x_{i-1})\right\|_{\mathrm{TV}}$$

and, for every $1 \le i < j \le n$,

$$a_i^j(x_i, y_i) = \left\|\mathscr{L}(X_j|X_i = x_i) - \mathscr{L}(X_j|X_i = y_i)\right\|_{\mathrm{TV}}.$$

We thus have, for every $1 \le i < j \le n$,

$$\gamma_i^j = \sup_{(x_i, y_i)\in\mathbb{R}^2} a_i^j(x_i, y_i),$$

and for every $2 \le i \le n$,

$$\alpha_i = \sup_{(x_{i-1}, y_{i-1})\in\mathbb{R}^2} b_i(x_{i-1}, y_{i-1}).$$

For every real-valued function v and for every probability measure K on $\mathbb{R}$, we denote

$$Kv = \int v\,dK.$$

According to this notation, for every $1 \le i < j \le n$,

$$\mathscr{L}(X_j|X_i = x_i) = K_{i+1}(\cdot|x_i)\cdots K_j(\cdot|\cdot).$$

Set

$$K_{i+1}(\cdot|x_i)\cdots K_j(\cdot|\cdot) = K_{i+1}^j(\cdot|x_i).$$

We want to bound uniformly $a_i^j(x_i, y_i)$. First, note that

$$\begin{aligned} &a_i^j(x_i, y_i) \\ &\quad = \left\|\int K_{i+2}^j(\cdot|x_{i+1})K_{i+1}(dx_{i+1}|x_i) - \int K_{i+2}^j(\cdot|y_{i+1})K_{i+1}(dy_{i+1}|y_i)\right\|_{\mathrm{TV}}. \end{aligned}$$

Define a coupling probability measure on $\mathbb{R}^2, \Pi(\cdot, \cdot|x_i, y_i)$, whose marginals are $K_{i+1}(\cdot|x_i)$ and $K_{i+1}(\cdot\,|y_i)$. Then,

$$a_i^j(x_i, y_i) = \left\|\iint K_{i+2}^j(\cdot|x_{i+1}) - K_{i+2}^j(\cdot|y_{i+1})\Pi(dx_{i+1}, dy_{i+1}|x_i, y_i)\right\|_{\mathrm{TV}}.$$

By convexity of the total variation norm $\|\cdot\|_{\mathrm{TV}}$,

$$a_i^j(x_i, y_i) \le \iint \left\|K_{i+2}^j(\cdot|x_{i+1}) - K_{i+2}^j(\cdot|y_{i+1})\right\|_{\mathrm{TV}}\Pi(dx_{i+1}, dy_{i+1}|x_i, y_i).$$

From the definition of γ_{i+1}^j, it follows that

$$a_i^j(x_i, y_i) \le (\gamma_{i+1}^j)^2 \iint \mathbf{1}_{x_{i+1}\neq y_{i+1}}\Pi(dx_{i+1}, dy_{i+1}|x_i, y_i).$$

Recall now the "coupling" definition of the variational distance between two measures of probability R and Q on $\mathbb{R}$,

$$\|Q - R\|_{\mathrm{TV}} = \min_{\Pi\in\mathscr{M}(Q, R)} \int \mathbf{1}_{x\neq y}\Pi(dx, dy),$$

where $\mathscr{M}(Q,R)$ is the set of all probability measures on $\mathbb{R}^2$ whose marginals are Q and R. Thanks to this definition, we choose the coupling probability measure $\Pi(\cdot,\cdot\,|x_i,y_i)$ in $\mathscr{M}(K_{i+1}(\cdot|x_i),K_{i+1}(\cdot\,|y_i))$ such that

$$b_{i+1}(x_i,y_i)=\iint \mathbf{1}_{x_{i+1}\neq y_{i+1}}\Pi(dx_{i+1},dy_{i+1}|x_i,y_i).$$

Consequently,

$$a_i^j(x_i,y_i)\le\left(\gamma_{i+1}^j\right)^2 b_{i+1}(x_i,y_i).$$

Thus, we obtain the following recurrence inequality, for every $1\le i<j\le n$,

$$\left(\gamma_i^j\right)^2\le\left(\gamma_{i+1}^j\right)^2\alpha_{i+1}.$$

Note that $\left(\gamma_{j-1}^j\right)^2=\alpha_j$ for every $2\le j\le n$. By induction over i, the preceding recurrence inequality immediately yields (2.9).

The second class of examples concerns Φ-mixing processes. For this class of examples, we refer to [5]. Consider X as a sample taken from a Φ-mixing random sequence $(X_i)_{i\in\mathbb{Z}}$. We briefly recall what is meant by this terminology. For any set C of integer, $C\subset\mathbb{Z}$, let $X_C=\{X_i,\,i\in C\}$, denote the C-marginals of the random process. $\mathscr{X}_C$ is the σ-algebra generated by X_C, and $|C|$ is the cardinal of C when C is finite. Moreover the usual distance between subsets A and B of $\mathbb{Z}$ will be denoted $d(A,B)$. To measure the Φ-dependence between two σ-algebras $\mathscr{X}_A$ and $\mathscr{X}_B$, definite

$$\Phi(\mathscr{X}_A,\mathscr{X}_B)=\sup\left\{\left|\mathbb{P}(V)-\frac{\mathbb{P}(U\cap V)}{\mathbb{P}(U)}\right|;\,U\in\mathscr{X}_A,\mathbb{P}(U)\neq 0,\,V\in\mathscr{X}_B\right\},$$

and for every integer k, u, v in $\mathbb{N}^*$,

$$\Phi_k(u,v)=\sup\{\Phi(\mathscr{X}_A,\mathscr{X}_B);d(A,B)\ge k,|A|\le u,|B|\le v\}.$$

We could observe that for each integer u and v, $\Phi_k(u,v)$ is nonincreasing with respect to k. The process $(X_i)_{i\in\mathbb{Z}}$ is said to be Φ-mixing, if for any integer, u, v,

$$\lim_{k\to\infty}\Phi_k(u,v)=0.$$

Note also that for every integer k, $\Phi_k(u,v)$ is nondecreasing with respect to u and v. We thus consider

$$\sup_{u\in\mathbb{N}^*}\sup_{v\in\mathbb{N}^*}\Phi_k(u,v)=\Phi_k.$$

Now let us present the relation between the coefficients Φ_k and the coefficients $\tilde{\gamma}_i^j$. We know that for all measures Q and R on a measurable space $(E,\mathscr{E})$, the variational distance can be defined as,

$$\|Q-R\|_{\mathrm{TV}}=\sup_{F\in\mathscr{E}}|Q(F)-R(F)|.$$

Recall the definition of the coefficient $\tilde{a}_j(y_1^i)$, for every $1\le i<j\le n$,

$$\tilde{a}_j(y_1^i)=\left\|\mathscr{L}(X_j^n|X_1^i=y_1^i)-\mathscr{L}(X_j^n)\right\|_{\mathrm{TV}}.$$

According to this definition, we see that

$$\operatorname*{ess\,sup}_{y_1^i \in \mathbb{R}^i,\, \mathscr{L}(X_1^i)} \tilde{a}_j(y_1^i) = \Phi(\mathscr{X}_{A(i)}, \mathscr{X}_{B(j)}),$$

where $A(i) = \{1, \dots, i\}$ and $B(j) = \{j, \dots, n\}$. Note that

$$d(A(i), B(j)) = j - i.$$

Consequently, form the definition (2.2) of the coefficient $\tilde{\gamma}_i^j$, it follows that

$$\left(\tilde{\gamma}_i^j\right)^2 \le 2\Phi_{j-i}.$$

Now, assume $(X_i)_{i\in\mathbb{Z}}$ is a Φ-mixing process for which the sequence (Φ_k) admits a geometrical decay; that is, for every k,

$$\Phi_k \le C\beta^k,$$

where C is some constant and β is a real number with $0 \le \beta < 1$. In this case, as for the previous examples, $\|\Gamma\|$ may also be bounded independently of n as

$$\|\Gamma\| \le \frac{\sqrt{2C}}{1 - \beta^{1/2}}.$$

More generally, we easily see that if $(X_i)_{i\in\mathbb{Z}}$ is a Φ-mixing process for which the sequence (Φ_k) satisfies

$$\sum_{k=1}^{\infty} \sqrt{\Phi_k} < \infty,$$

then $\|\Gamma\|$ may be bounded independently of the size n of the sample X as

$$\|\Gamma\| \le \sum_{k=1}^{\infty} \sqrt{\Phi_k}.$$

There are probably other examples of samples X for which $\|\Gamma\|$ may be bounded independently of n, but there are not developed in this paper. We now present the central theorem of this paper.

This theorem is an improvement of the theorem by Marton [8]. For every measure of probability Q and R on $\mathbb{R}^n$, let $\mathscr{M}(Q, R)$ denote the set of all probability measures on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals Q and R. Define

$$d_2(Q, R) = \inf_{\Pi \in \mathscr{M}(Q, R)} \sup_{\alpha} \iint \sum_{i=1}^{n} \alpha_i(y) \mathbf{1}_{x_i \ne y_i}\, d\Pi(x, y), \tag{2.10}$$

where $\sup_\alpha$ is over all vectors of positive functions $\alpha = (\alpha_1, \dots, \alpha_n)$, with

$$\int \sum_{i=1}^{n} \alpha_i^2(y)\, dR(y) \le 1.$$

This definition is due to Marton. Let us note that Marton rather uses the normalized distance $\bar{d}_2 = d_2/\sqrt{n}$. It is clear that $d_2(Q, R)$ is not symmetric and that $d_2(Q, Q) = 0$. Marton proved that d_2 satisfies a triangular inequality (see [9]). Therefore, $d_2(Q, R)$ is quite a distance between the measures Q and R. Another expression for the distance $d_2(Q, R)$ is the following (see [8]):

$$d_2(Q, R) = \inf_{\Pi \in \mathscr{M}(Q, R)} \left(\int \sum_{i=1}^{n} Pr^2(X_i \neq y_i | Y_i = y_i)\, dR(y) \right)^{1/2},$$

where (X, Y) denotes a pair of random variables taking values in $\mathbb{R}^n \times \mathbb{R}^n$, and with law Π.

Let $X = (X_1, \dots, X_n)$ be a sample of random variables taking values in $\mathbb{R}^n$. For example, X is one of the previous examples. Let Γ be its corresponding matrix of mixing coefficients defined in (2.1) or (2.2). As defined previously, P denotes the law of the sample X.

THEOREM 1. *For every probability measures Q on $\mathbb{R}^n$ with Radon–Nikodym derivative dQ/dP with respect to the measure P,*

$$d_2(Q, P) \leq \|\Gamma\| \sqrt{2\, \mathrm{Ent}_P\left(\frac{dQ}{dP}\right)}. \tag{2.11}$$

Furthermore,

$$d_2(P, Q) \leq \|\Gamma\| \sqrt{2\, \mathrm{Ent}_P\left(\frac{dQ}{dP}\right)}. \tag{2.12}$$

As a consequence of this main result, we present a few corollaries of interest. $X = (X_1, \dots, X_n)$ is a sample of bounded random variables. It will be convenient to assume that each X_i takes values in $[0, 1]$. The results easily extend to arbitrary bounded random variables as for inequality (2.22). The support of P is also on $[0, 1]^n$. We just have to change the definition of the coefficients of the matrix Γ, replacing the set $\mathbb{R}$ by the set $[0, 1]$ in (2.1) and (2.2).

The first corollary concerns log-Sobolev inequalities for the measure P and convex or concave smooth functions on $[0, 1]^n$. As already mentioned in the introduction, this corollary extends Theorem 1.2 of [7] which concerns log-Sobolev inequalities for product measures on $[0, 1]^n$ and for separately convex smooth functions.

COROLLARY 1. *For any smooth convex function $f\colon [0, 1]^n \to \mathbb{R}$,*

$$\mathrm{Ent}_P\left(e^f\right) \leq 2\|\Gamma\|^2 \int |\nabla f|^2 e^f\, dP. \tag{2.13}$$

For any smooth concave function $f\colon [0, 1]^n \to \mathbb{R}$,

$$\mathrm{Ent}_P\left(e^f\right) \leq 2\|\Gamma\|^2 \int |\nabla f|^2\, dP \int e^f\, dP \tag{2.14}$$

(where ∇f is the usual gradient of f on $\mathbb{R}^n$ and $|\nabla f|$ denotes its Euclidean length).

Theorem 1 yields lower bounds for the informational divergence between measures. Conversely, we bound the entropy of functions in Corollary 1. We could say that the log-Sobolev inequalities (2.13) and (2.14) are the dual expressions of the information inequalities (2.11) and (2.12).

PROOF. On the basis of Theorem 1, the proof of Corollary 1 is quite simple. Our aim is to bound efficiently $\mathrm{Ent}_P(e^f)$ with the usual gradient of f. By Jensen's inequality, for any function f,

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq \int f(y)\frac{e^{f(y)}}{\mathrm{E}_P(e^f)}\,dP(y) - \int f(x)\,dP(x).$$

Let P^f be the probability measure on $[0,1]^n$ whose density is $e^f/\mathrm{E}_P(e^f)$ with respect to the measure P. Let Π be a probability measure on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals P and P^f. Then

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq \iint f(y) - f(x)\,d\Pi(x,y).$$

Let f be a convex function on $[0,1]^n$. For every x and y in $[0,1]^n$, we can bound $f(y)-f(x)$ independently of $\nabla f(x)$. More precisely,

$$f(y_1,\dots,y_n) - f(x_1,\dots,x_n) \leq \sum_{j=1}^n |\partial_j f(y_1,\dots,y_n)|\,|y_j - x_j|.$$

For every y_j, x_j in $[0, 1]$, $|y_j - x_j| \leq \mathbf{1}_{y_j\neq x_j}$, so that

$$f(y_1,\dots,y_n) - f(x_1,\dots,x_n) \leq \sum_{j=1}^n |\partial_j f(y_1,\dots,y_n)|\mathbf{1}_{y_j\neq x_j}. \tag{2.15}$$

Let us recall that $\mathscr{M}(P,P^f)$ denotes the set of all probability measures on $\mathbb{R}^n\times\mathbb{R}^n$ with marginals P and P^f. Therefore, for every probability measure Π in $\mathscr{M}(P,P^f)$,

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq \sum_{j=1}^n \iint |\partial_j f(y)|\mathbf{1}_{y_j\neq x_j}\Pi(x,y).$$

Similarly, if f is a concave function on $[0,1]^n$, for every x and y in $[0,1]^n$, we can bound $f(y)-f(x)$ independently of $\nabla f(y)$.

$$f(y_1,\dots,y_n) - f(x_1,\dots,x_n) \leq \sum_{j=1}^n \big|\partial_j f(x_1,\dots,x_n)\big|\mathbf{1}_{y_j\neq x_j}. \tag{2.16}$$

Therefore, we get in this case

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq \sum_{j=1}^n \iint \big|\partial_j f(x)\big|\mathbf{1}_{y_j\neq x_j}, \Pi(x,y).$$

According to the definitions of $d_2(P, P^f)$ and $d_2(P^f, P)$, by the Cauchy–Schwarz inequality, for every convex function f on $[0, 1]^n$,

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq d_2(P, P^f)\left(\int \sum_{j=1}^n \left|\partial_j f(y)\right|^2 dP^f(y)\right)^{1/2}.$$

Similarly, for every concave function f on $[0, 1]^n$,

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq d_2(P^f, P)\left(\int \sum_{j=1}^n \left|\partial_j f(x)\right|^2 dP(x)\right)^{1/2}.$$

Apply then the results of Theorem 1 to $d_2(P, P^f)$ and $d_2(P^f, P)$. Since

$$\frac{dP^f}{dP} = \frac{e^f}{\mathrm{E}_P(e^f)},$$

we get, for every convex function f on $[0, 1]^n$,

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq \|\Gamma\|\left(2\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)}\right)^{1/2}\left(\int |\nabla f|^2 \frac{e^f}{\mathrm{E}_P(e^f)}\, dP\right)^{1/2}.$$

Similarly, for every concave function f on $[0, 1]^n$,

$$\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)} \leq \|\Gamma\|\left(2\frac{\mathrm{Ent}_P(e^f)}{\mathrm{E}_P(e^f)}\right)^{1/2}\left(\int |\nabla f|^2\, dP\right)^{1/2}.$$

The proof is thus complete. □

A direct application of Corollary 1 is Poincaré or spectral gap inequalities for convex or concave functions. Let f be a convex function. For any ε positive, apply (2.13) of Theorem.1, to εf. A Taylor's expansion of the second order in ε in (2.13) yields the following corollary.

COROLLARY 2. *For any smooth convex real function f on* $[0, 1]^n$,

$$\int f^2\, dP - \left(\int f\, dP\right)^2 \leq 2\|\Gamma\|^2 \int |\nabla f|^2\, dP. \tag{2.17}$$

Note that this inequality has been proved with a better constant in the independent case ($\|\Gamma\| = 1$) in [1] and [7].

We now present new concentration inequalities. Obviously, using the classical method developed by Ledoux (see also [7]), we easily derive deviation inequalities from the log-Sobolev inequalities (2.13) and (2.14). However, we get the constant 8 instead of the better constant 2 in the deviation inequalities (2.18) and (2.19). The way to obtain the optimal constant 2 is to adapt a proof by Bobkov and Götze [2] to the information inequalities of Theorem 1.

COROLLARY 3. *For any smooth convex function f on $[0, 1]^n$ satisfying $|\nabla f| \leq 1$ P-almost everywhere, for every $t \geq 0$,*

$$P(f \geq \mathrm{E}_P(f) + t) \leq \exp\left(-\frac{t^2}{2\|\Gamma\|^2}\right). \tag{2.18}$$

For any smooth concave function f on $[0,1]^n$ satisfying $\int |\nabla f|^2\, dP \leq 1$, for every $t \geq 0$,

$$P(f \geq \mathrm{E}_P(f) + t) \leq \exp\left(-\frac{t^2}{2\|\Gamma\|^2}\right). \tag{2.19}$$

These deviation inequalities are of particular interest if $\|\Gamma\|$is bounded independently of the size n of the sample X. Let us note that for the same exponential deviation inequality (2.19) or (2.18), the condition on the gradient is stronger for convex functions than for concave functions. This is rather intuitive on the graph of a concave function and its mean. Equation (2.19) thus improves some aspects of the results in [6], recalled further in this paper [see (2.21)]. On the other hand, we do not deal with separately convex functions as in [7]. It might be of interest to find the information inequalities that would cover this class of functions.

Corollary 3 yields a concentration inequality for convex (or concave) Lipschitz functions. Let f be a convex Lipschitz function on $\mathbb{R}^n$ with Lipschitz constant

$$\|f\|_{\mathrm{Lip}} \leq 1.$$

Let $P_\varepsilon f$ be the convolution product of f with a Gaussian kernel, for every ε positive, for every x in $\mathbb{R}^n$,

$$P_\varepsilon f(x) = \int f(y) \exp\left(-\frac{|x-y|^2}{2\varepsilon}\right)\frac{d\lambda(y)}{\sqrt{2\pi\varepsilon}} = \mathrm{E}(f(x + \sqrt{\varepsilon}B)),$$

where λ is the Lebesgue measure on $\mathbb{R}^n$, and B denotes a Gaussian variable on $\mathbb{R}^n$ whose law is $N(0, I)$. Clearly $P_\varepsilon f$ is a convex function on $\mathbb{R}^n$. Since $\|f\|_{\mathrm{Lip}} \leq 1$, for every x in $\mathbb{R}^n$,

$$|P_\varepsilon f(x) - f(x)| \leq \sqrt{\varepsilon}\,\mathrm{E}\,|B|.$$

Therefore, for every x in $\mathbb{R}^n$, $P_\varepsilon f(x)$ converges to $f(x)$ as ε tends to 0. Moreover, by Rademacher's theorem,

$$|\nabla f| \leq 1 \quad \lambda\text{-almost everywhere.}$$

Consequently,

$$|\nabla P_\varepsilon f| \leq 1 \quad \text{everywhere,}$$

since

$$\nabla P_\varepsilon f(x) = \mathrm{E}(\nabla f(x + \sqrt{\varepsilon}B)).$$

We then apply (2.18) to $P_\varepsilon f$ and (2.19) to $-P_\varepsilon f$. This yields the following result, for every $t \geq 0$,

$$P(|P_\varepsilon f - \mathrm{E}_P(P_\varepsilon f)| \geq t) \leq 2\exp\biggl(-\frac{t^2}{2\|\Gamma\|^2}\biggr).$$

Since $P_\varepsilon f(x)$ converges to $f(x)$ everywhere as ε tends to 0, we get the following corollary.

COROLLARY 4. *For any convex Lipschitz function f on $[0,1]^n$ with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leq 1$, and for every $t \geq 0$,*

$$P(|f - \mathrm{E}_P(f)| \geq t) \leq 2\exp\biggl(-\frac{t^2}{2\|\Gamma\|^2}\biggr). \tag{2.20}$$

If P is a product measure $\mu_1 \otimes \cdots \otimes \mu_n$ on $[0,1]^n$, $\|\Gamma\| = 1$, the latter inequality (2.20) is the analogue of Talagrand's deviation inequality with the median M instead of the mean (See also [15], [16], [17]). Talagrand showed that for every convex Lipschitz function f, with $\|f\|_{\mathrm{Lip}} \leq 1$, for every $t \geq 0$,

$$P(|f - M| \geq t) \leq 4\exp(-t^2/4). \tag{2.21}$$

Marton extends this result to contracting Markov chains. Talagrand and Marton first prove the concentration of measure phenomenon in terms of sets (see Theorem 6.1, [16]). (2.21) follows by considering the set $\{f \leq M\}$ (see Theorem 6.6, [16]). Actually, concentration inequalities around the mean or the median are equivalent up to numerical constants (see, e.g., [14]). Let us briefly sketch the argument. Corollary 4 indicates that, for a convex Lipschitz function with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leq 1$, if $t > 4\|\Gamma\|\sqrt{\log 2}$,

$$P(|f - \mathrm{E}_P(f)| \leq t) > \tfrac{1}{2}.$$

Therefore, the definition of the median implies that

$$|M - \mathrm{E}_P(f)| \leq \sqrt{2}.$$

Thus, from (2.20), we get that, for every $u \geq 0$,

$$P(|f - M| \geq u) \leq P(|f - \mathrm{E}_P(f)| \geq u - \sqrt{2}) \leq 2\exp\biggl(-\frac{(u-\sqrt{2})^2}{2\|\Gamma\|^2}\biggr).$$

Hence, since $\|\Gamma\| \geq 1$, for every $u \geq 0$,

$$P(|f - M| \geq u) \leq 6\exp\biggl(-\frac{u^2}{4\|\Gamma\|^2}\biggr).$$

Corollary 4 of course extends to probability measures P on $[a,b]^n$. Assume P is the distribution of a sample $X = (X_1, \ldots, X_n)$ of random variables on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$. Each random variable X_i takes values in

$[a, b]$. By a simple scaling, we get from Corollary 4 that for any convex Lipschitz function f on $\mathbb{R}^n$, with Lipschitz constant $\|f\|_{\mathrm{Lip}} \le 1$, for every $t \ge 0$,

$$P(|f - \mathrm{E}_P(f)| \ge t) \le 2\exp\left(-\frac{t^2}{2(b-a)^2\|\Gamma\|^2}\right). \tag{2.22}$$

Let us also recall one typical application of this deviation inequality to norms of random series. For $1 \le i \le n$, let Z_i be random variables on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ with $|Z_i| \le 1$. Γ denotes its triangular matrix of mixing coefficients. For $1 \le i \le n$, let b_i be vectors in some arbitrary Banach space E with norm $\|\cdot\|$. Then, for every $t \ge 0$,

$$\mathbb{P}\left(\left|\left\|\sum_{i=1}^n Z_i b_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n Z_i b_i\right\|\right| \ge t\right) \le \exp\left(-\frac{t^2}{8\sigma^2\|\Gamma\|^2}\right), \tag{2.23}$$

where

$$\sigma^2 = \sup_{\|\xi\| \le 1} \sum_{i=1}^n \langle \xi, b_i\rangle^2.$$

We now turn to the proof of Corollary 3. Instead of using the method suggested by Marton, dealing with a geometric description of concentration, we prefer to follow the functional approach of [7]. Our approach is inspired by [2]. The following proof of Corollary 3 is an adaptation of the proof of Theorem 3.1 of [2] to the particular case of a nonsymmetric d_2-distance between probability measures on $\mathbb{R}^n$. The proof is based on the relation between the information inequalities of Theorem 1 and exponential integrability.

PROOF OF COROLLARY 3. Let f be a convex function on $[0, 1]^n$. Let Q be a measure on $[0, 1]^n$ with Radon–Nikodym derivative $dQ/dP = g$ with respect to the measure P. For every measure Π in $\mathscr{M}(P, Q)$, that is, for every measure Π on $\mathbb{R}^n \times \mathbb{R}^n$, whose marginals are Q and P,

$$\int f(y)\,dQ(y) - \int f(x)\,dP = \iint (f(y) - f(x))\,d\Pi(x, y).$$

As already mentioned in the proof of Corollary 1, if f is a convex function on $[0, 1]^n$, we can bound $f(y) - f(x)$ independently of $\nabla f(x)$. Namely, for every x and y in $[0, 1]^n$,

$$f(y_1, \dots, y_n) - f(x_1, \dots, x_n) \le \sum_{j=1}^n |\partial_j f(y)|\,\mathbf{1}_{y_j \ne x_j}.$$

Therefore, for every measure Π in $\mathscr{M}(P, Q)$,

$$\int f(y)\,dQ(y) - \int f(x)\,dP(x) \le \iint \sum_{j=1}^n |\partial_j f(y)| \mathbf{1}_{y_j \ne x_j}\,d\Pi(x, y).$$

The assumption that $|\nabla f| \le 1$ P-almost everywhere is still true Q-almost everywhere. Therefore,

$$\int \sum_{j=1}^{n} |\partial_j f|^2 \, dQ \le 1.$$

Finally, according to the definition of $d_2(P, Q)$ (2.10), we get that,

$$\int f \, dQ - \int f \, dP \le d_2(P, Q). \tag{2.24}$$

Similarly, if f is a concave function on $[0, 1]^n$, for every x and y in $[0, 1]^n$, we bound $f(y) - f(x)$ independently $\nabla f(y)$. Therefore, we get in this case,

$$\int f(y) \, dQ(y) - \int f(x) \, dP(x) \le \iint \sum_{j=1}^{n} |\partial_j f(x)| \mathbf{1}_{y_j \neq x_j} \, d\Pi(x, y).$$

Under the assumption that

$$\int \sum_{j=1}^{n} |\partial_j f|^2 \, dP \le 1,$$

it follows that

$$\int f \, dQ - \int f \, dP \le d_2(Q, P). \tag{2.25}$$

Assume now that f is either a convex function, or a concave function, satisfying the assumption of Corollary 3. Applying the results of Theorem 1, (2.11) or (2.12), we get from (2.24) or (2.25) that

$$\int f \, dQ - \int f \, dP \le \sqrt{2\|\Gamma\|^2 \operatorname{Ent}_P\left(\frac{dQ}{dP}\right)}.$$

That is,

$$\int fg \, dP - \int f \, dP \le \sqrt{2\|\Gamma\|^2 \operatorname{Ent}_P(g)}.$$

We then use the following variational equality:

$$\sqrt{2\|\Gamma\|^2 \operatorname{Ent}_P(g)} = \inf_{\lambda > 0} \left(\frac{\|\Gamma\|^2 \lambda}{2} + \frac{1}{\lambda} \operatorname{Ent}_P(g) \right).$$

Thus, for every $\lambda > 0$,

$$\int (f - \mathrm{E}_P(f)) g \, dP \le \frac{\|\Gamma\|^2 \lambda}{2} + \frac{1}{\lambda} \operatorname{Ent}_P(g).$$

In other words, for every $\lambda > 0$,

$$\int \left(\lambda (f - \mathrm{E}_P(f)) - \frac{\|\Gamma\|^2 \lambda^2}{2} \right) g \, dP \le \operatorname{Ent}_P(g).$$

Then, choosing $g = e^l / \mathrm{E}_P\left(e^l\right)$ where

$$l = \lambda(f - \mathrm{E}_P(f)) - \frac{\|\Gamma\|^2\lambda^2}{2},$$

it follows that for every $\lambda \geq 0$,

$$\mathrm{E}_P\big(e^{\lambda(f-\mathrm{E}_P(f))}\big) \leq \exp\left(\frac{\|\Gamma\|^2\lambda^2}{2}\right).$$

By Chebyshev's inequality, for every $\lambda \geq 0$, $t \geq 0$,

$$P(f - \mathrm{E}_P(f) \geq t) \leq \exp\left(-\lambda t + \frac{\|\Gamma\|^2\lambda^2}{2}\right).$$

Optimizing in λ proves the deviation inequalities (2.18) and (2.19) of Corollary 3. □

We now turn to the (some what lengthy) proof of Theorem 1. To better explain the idea, let us first outline the scheme of the proof. We first assume that P admits a strictly positive density g with respect to a product measure $\mu_1 \otimes \cdots \otimes \mu_n$ on $[0,1]^n$. This assumption is not restrictive. Indeed, consider the case of a nonnegative density $\tilde{g}$. Let then

$$g = \tilde{g}\mathbf{1}_{\tilde{g}>0}.$$

Here g is a strictly positive measurable function. So we can consider the probability whose density is g with respect to $\mu_1 \otimes \cdots \otimes \mu_n$ on $[0,1]^n$. We then apply Theorem 1 to this measure. Noting that $\mathbf{1}_{\tilde{g}=0}$ is a measurable function, we easily extend the results of Theorem 1 to the case of a nonnegative density $\tilde{g}$.

Let Q be a probability measure on $\mathbb{R}^n$, with Radon–Nikodym derivative dQ/dP with respect to P. Let α be a vector of positive functions $\alpha = (\alpha_1, \ldots, \alpha_n)$, with

$$\int \sum_{i=1}^n \alpha_i^2(y)\, dQ(y) \leq 1.$$

Let β be a vector of positive functions $\beta = (\beta_1, \ldots, \beta_n)$, with

$$\int \sum_{i=1}^n \beta_i^2(x)\, dP(x) \leq 1.$$

The key of the proof is to find a good measure Π with marginals Q and P to bound efficiently the two following expressions independently of α or β. Precisely, we will construct a measure Π such that, for every α and β with the above conditions,

$$\iint \sum_{i=1}^n \alpha_i(y)\mathbf{1}_{x_i \neq y_i}\, d\Pi(x, y) \leq \|\Gamma\| \sqrt{2\, \mathrm{Ent}_P\left(\frac{dQ}{dP}\right)}$$

and

$$\iint \sum_{i=1}^{n} \beta_i(x) \mathbf{1}_{x_i \neq y_i} \, d\Pi(x, y) \leq \|\Gamma\| \sqrt{2 \operatorname{Ent}_P \left(\frac{dQ}{dP} \right)}.$$

To this task, we introduce conditioning notation. If g is a strictly positive density, we can write,

$$g(x_1, \ldots, x_n) = g_n(x_n | x_1, \ldots, x_{n-1}) \cdots g_2(x_2 | x_1) g_1(x_1),$$

where for $1 \leq j \leq n$,

$$g_j(x_j | x_1, \ldots, x_{j-1}) = \frac{\int g(x_1, \ldots, x_j, z_{j+1}, \ldots, z_n) \mu_{j+1}(dz_{j+1}) \cdots \mu_n(dz_n)}{\int g(x_1, \ldots, x_{j-1}, z_j, \ldots, z_n) \mu_j(dz_j) \cdots \mu_n(dz_n)}.$$

For $1 \leq j \leq n$, we denote by $G_j(\cdot | x_1, \ldots, x_{j-1})$ the probability measure whose density is $g_j(\cdot | x_1, \ldots, x_{j-1})$ with respect to the measure μ_j,

$$G_j(dx_j | x_1, \ldots, x_{j-1}) = g_j(x_j | x_1, \ldots, x_{j-1}) \mu_j(dx_j).$$

Let h denotes the density of the measure Q with respect to the product measure $\mu_1 \otimes \cdots \otimes \mu_n$ on $\mathbb{R}^n$,

$$h = \frac{dQ}{dP} g.$$

Similarly for the density h,

$$h(y_1, \ldots, y_n) = h_n(y_n | y_1, \ldots, y_{n-1}) \cdots h_2(y_2 | y_1) h_1(y_1)$$

with

$$h_j(y_j | y_1, \ldots, y_{j-1}) = \frac{\int h(y_1, \ldots, y_j, z_{j+1}, \ldots, z_n) \mu_{j+1}(dz_{j+1}) \cdots \mu_n(dz_n)}{\int h(y_1, \ldots, y_{j-1}, z_j, \ldots, z_n) \mu_j(dz_j) \cdots \mu_n(dz_n)}.$$

We set similarly

$$H_j(dx_j | x_1, \ldots, x_{j-1}) = h_j(x_j | x_1, \ldots, x_{j-1}) \mu_j(dx_j).$$

To clarify all the proof, we need some additional conditioning notation. For every $1 \leq i < j \leq k \leq n$, let

$$\begin{aligned} &h_j^k(y_j, \ldots, y_k | y_1, \ldots, y_i) \\ &\quad = \int \cdots \int h(y_1, \ldots, y_i, z_{i+1}, \ldots, z_{j-1}, y_j, \ldots, y_k, z_{k+1}, \ldots, z_n) \\ &\qquad\qquad \times \mu_{i+1}(dz_{i+1}) \cdots \mu_{j-1}(dz_{j-1}) \mu_{k+1}(dz_{k+1}) \cdots \mu_n(dz_n). \end{aligned}$$

To simplify the notation, $H_j^k(\cdot, \ldots, \cdot | y_1, \ldots, y_i)$ will denote the probability measure whose density is $h_j^k(\cdot, \ldots, \cdot | y_1, \ldots, y_i)$, that is,

$$H_j^k(dy_j, \ldots, dy_k | y_1, \ldots, y_i) = h_j^k(y_j, \ldots, y_k | y_1, \ldots, y_i) \mu_j(dy_j) \cdots \mu_k(dy_k).$$

Similarly, with the same definitions for g_j^k and G_j^k,

$$G_j^k(dy_j, \dots, dy_k | y_1, \dots, y_i) = g_j^k(y_j, \dots, y_k | y_1, \dots, y_i)\mu_j(dy_j) \cdots \mu_k(dy_k).$$

Moreover, we sometimes write y_1^j for $(y_1, \dots, y_j)$, $1 \le j \le n$. Let us note that, for $1 \le j \le n$,

$$h_1^{j-1}(y_1, \dots, y_{j-1} = h_{j-1}(y_{j-1} | y_{j-2}, \dots, y_1) \cdots h_2(y_2 | y_1) h_1(y_1)$$

and

$$P = G_1^n, \qquad P^f = H_1^n.$$

With these notation, we set, for $1 \le i \le n$,

$$E_i = \int \mathrm{Ent}_{G_i(\cdot | y_1, \dots, y_{i-1})}\left(\frac{h_i(\cdot | y_1, \dots, y_{i-1})}{g_i(\cdot | y_1, \dots, y_{i-1})}\right) H_1^{i-1}(dy_1, \dots, dy_{i-1}).$$

Let us recall the well-known tensorization property of entropy.

LEMMA 1.

$$\sum_{i=1}^n E_i = \mathrm{Ent}_P\left(\frac{h}{g}\right) = \mathrm{Ent}_P\left(\frac{dQ}{dP}\right). \tag{2.26}$$

Together with Lemma 2 below, this property is one main argument of the proof of Theorem 1.

PROOF. We have

$$\mathrm{Ent}_P\left(\frac{h}{g}\right) = \int \frac{h}{g} \log \frac{h}{g}\, dP.$$

Since

$$\frac{h(y_1, \dots, y_n)}{g(y_1, \dots, y_n)} = \frac{h_n(y_n | y_1, \dots, y_{n-1})}{g_n(y_n | y_1, \dots, y_{n-1})} \cdots \frac{h_1(y_1)}{g_1(y_1)},$$

it follows that

$$\mathrm{Ent}_P\left(\frac{h}{g}\right) = \sum_{i=1}^n \int \cdots \int \log \frac{h_i(y_i | y_1, \dots, y_{i-1})}{g_i(y_i | y_1, \dots, y_{i-1})} H_1^n(dy_1, \dots, dy_n).$$

Integrating, this yields

$$\begin{aligned} \mathrm{Ent}_P\left(\frac{h}{g}\right) = \sum_{i=1}^n \int \cdots \int \log \frac{h_i(y_i | y_1, \dots, y_{i-1})}{g_i(y_i | y_1, \dots, y_{i-1})} \\ \times H_i(dy_i | y_1, \dots, y_{i-1}) H_1^{i-1}(dy_1, \dots, dy_{i-1}). \end{aligned} \tag{2.27}$$

According to the definition of entropy,

$$\operatorname{Ent}_{G_i(\cdot|y_1,\ldots,y_{i-1})}\left(\frac{h_i(\cdot|y_1,\ldots,y_{i-1})}{g_i(\cdot|y_1,\ldots,y_{i-1})}\right)$$
$$=\int\log\frac{h_i(y_i|y_1,\ldots,y_{i-1})}{g_i(y_i|y_1,\ldots,y_{i-1})}H_i(dy_i|y_1,\ldots,y_{i-1}).$$

Consequently, with the definition of E_i, we see that (2.27) is equivalent to (2.26). □

For $1\le j\le n$, consider

$$\Delta_j=\int\alpha_j(y)^2\,dQ(y)$$

and

$$\tilde{\Delta}_j=\int\beta_j(x)^2\,dP(x).$$

To be more precise, to prove Theorem 1, we will construct a measure Π such that, for every $1\le j\le n$,

$$\iint\alpha_j(y)\mathbf{1}_{y_j\neq x_j}\,d\Pi(y,x)\le\sum_{i=1}^{j}\gamma_i^j(2E_i)^{1/2}(\Delta_j)^{1/2}\tag{2.28}$$

and

$$\iint\beta_j(x)\mathbf{1}_{y_j\neq x_j}\,d\Pi(y,x)\le\sum_{i=1}^{j}\gamma_i^j(2E_i)^{1/2}(\tilde{\Delta}_j)^{1/2}.\tag{2.29}$$

Then, according to the definition of the usual operator norm of the matrix Γ with respect to the Euclidean topology, it follows that

$$\iint\sum_{i=1}^{n}\alpha_i(y)\mathbf{1}_{x_i\neq y_i}\,d\Pi(x,y)\le\|\Gamma\|\left(2\sum_{i=1}^{n}E_i\right)^{1/2}\left(\sum_{j=1}^{n}\Delta_j\right)^{1/2}$$

and

$$\iint\sum_{i=1}^{n}\beta_i(x)\mathbf{1}_{x_i\neq y_i}\,d\Pi(x,y)\le\|\Gamma\|\left(2\sum_{i=1}^{n}E_i\right)^{1/2}\left(\sum_{j=1}^{n}\tilde{\Delta}_j\right)^{1/2}.$$

By the definitions of Δ_j and $\tilde{\Delta}_j$,

$$\sum_{j=1}^{n}\Delta_j\le 1$$

and

$$\sum_{j=1}^{n}\tilde{\Delta}_j\le 1.$$

The information inequalities (2.11) and (2.12) of Theorem 1 will then follow from Lemma 1.

So, to prove Theorem 1, we just have to show (2.28) and (2.29). Before considering the general case, it is of interest to see the case of dimension one, $n = 1$. In this proof, we present Lemma 2 which is at the center of the proof of Theorem 1. Then, to extend our approach to any dimension, we use a result of Fiebig in [6] recalled in Proposition 2.

If $n = 1$, we want to construct a measure Π on $\mathbb{R} \times \mathbb{R}$ with marginals P and Q. Let Π be the probability whose density l_1 with respect to $\mu_1(dy_1) \otimes \mu_1(dx_1)$ is defined by

$$\begin{aligned} l_1(x_1, y_1) &= \mathbf{1}_{x_1 = y_1} \min(h(y_1), g(x_1)) \\ &\qquad + \mathbf{1}_{x_1 \neq y_1} \frac{[h(y_1) - g(y_1)]_+ [g(x_1) - h(x_1)]_+}{\|Q - P\|_{\mathrm{TV}}}, \end{aligned}$$

where $[\alpha]_+$ denotes the positive part of the real number α. Integrating one of the variables, it is clear that the marginals of Π are Q and P. With this definition, we have,

$$\begin{aligned} &\iint \alpha_1(y_1) \mathbf{1}_{y_1 \neq x_1} \Pi(dy_1, dx_1) \\ &\quad = \iint \alpha_1(y_1) \mathbf{1}_{y_1 \neq x_1} \frac{[h(y_1) - g(y_1)]_+ [g(x_1) - h(x_1)]_+}{\|Q - P\|_{\mathrm{TV}}} \mu_1(dy_1) \mu_1(dx_1). \end{aligned}$$

We know that

$$\int [g(x_1) - h(x_1)]_+ \, \mu_1(dx_1) = \|Q - P\|_{\mathrm{TV}}.$$

Therefore, integrating with respect to the variable x_1, it follows that

$$\iint \alpha_1(y_1) \mathbf{1}_{y_1 \neq x_1} \Pi(dy_1, dx_1) = \int \alpha_1(y_1) [h(y_1) - g(y_1)]_+ \, \mu_1(dy_1).$$

Since

$$\int \alpha_1(y_1)^2 \, dQ(y_1) \leq 1,$$

by the Cauchy–Schwarz inequality, we get that

$$\iint \alpha_1(y_1) \mathbf{1}_{y_1 \neq x_1} \Pi(dy_1, dx_1) \leq \left(\int \left[1 - \frac{g(y_1)}{h(y_1)} \right]_+^2 h(y_1) \, \mu_1(dy_1) \right)^{1/2}. \tag{2.30}$$

Similarly, with the same definition for the measure Π, we have

$$\iint \beta(x_1) \mathbf{1}_{y_1 \neq x_1} \Pi(dy_1, dx_1) \leq \left(\int \left[1 - \frac{h(x_1)}{g(x_1)} \right]_+^2 g(x_1) \, \mu_1(dx_1) \right)^{1/2}. \tag{2.31}$$

Finally, to end the proof in the case $n = 1$, we just have to apply the following lemma.

LEMMA 2. *For every probability measures R and Q with density r and q with respect to a measure ν, define*

$$d_\nu(r|q) = \left(\int \left[1 - \frac{r}{q}\right]_+^2 q\, d\nu\right)^{1/2}.$$

Then, we have

$$d_\nu^2(r|q) + d_\nu^2(q|r) \le 2\,\mathrm{Ent}_R\left(\frac{q}{r}\right).$$

Consequently,

$$d_\nu(r|q) \le \left(2\,\mathrm{Ent}_R\left(\frac{q}{r}\right)\right)^{1/2} \tag{2.32}$$

and

$$d_\nu(q|r) \le \left(2\,\mathrm{Ent}_R\left(\frac{q}{r}\right)\right)^{1/2}. \tag{2.33}$$

This result is an improvement of Lemma 3.2 of [8]. Indeed Marton proves (2.32) and (2.33) without giving the upper symmetric version. Moreover, we will present a simpler proof of it. However, let us note that the proof of Theorem 1 will only use the nonsymmetric inequalities (2.32) and (2.33).

Note that for $n = 1$, we exactly have, with our notation,

$$d_\nu(r|q) = d_2(R, Q) = \inf_{\Pi \in \mathscr{M}(R,\, Q)} \left(\int Pr(Z \neq y_1 | Y = y_1)^2\, dQ(y_1)\right)^{1/2},$$

where (Z, Y) is a pair of a random variables taking values in $\mathbb{R} \times \mathbb{R}$, with law Π.

PROOF. Let $u = q/r$. We have

$$\mathrm{Ent}_R(u) = \int (u \log u - u + 1) r\, d\nu. \tag{2.34}$$

Let

$$\Psi(u) = u \log u - u + 1$$

and

$$\Phi(u) = \frac{\Psi(u)}{u}.$$

An elementary study of the functions Ψ and Φ shows that, for every $0 \le u \le 1$,

$$\Psi(u) \ge \tfrac{1}{2}(1 - u)^2,$$

whereas for $u \ge 1$,

$$\Phi(u) \ge \frac{1}{2}\left(1 - \frac{1}{u}\right)^2.$$

Since

$$u\log u - u + 1 = \Psi(u)\mathbf{1}_{u\le 1} + u\Phi(u)\mathbf{1}_{u\ge 1},$$

it follows that

$$u\log u - u + 1 \ge \frac{1}{2}\big[1-u\big]_+^2 + u\frac{1}{2}\Big[1-\frac{1}{u}\Big]_+^2.$$

Making use of this inequality in (2.34) ends the proof of Lemma 2. □

PROOF OF THEOREM 1. We want to generalize the preceding argument to any dimension n. We just have to construct a measure Π on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals P and Q satisfying the inequalities (2.28)and (2.29). In fact, we will construct a measure $\overline{\Pi}$ on

$$\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n-1} \times \cdots \times \mathbb{R}$$

with marginals P and Q. The construction of $\overline{\Pi}$ is not as simple as in the case $n = 1$. Before giving the expression of $\overline{\Pi}$, we will present step by step the structure of dependence between random variables

$$(Y_1, \dots, Y_n), \Big(X_1^{(1)}, \dots, X_n^{(1)}\Big), \Big(X_2^{(2)}, \dots, X_n^{(2)}\Big), \dots, \Big(X_{n-1}^{(n-1)}, X_n^{(n-1)}\Big), X_n^{(n)}$$

taking values in

$$\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n-1} \times \cdots \times \mathbb{R},$$

with law $\overline{\Pi}$ on $(\Omega, \mathscr{A}, \mathbb{P})$. To simplify the notation, for every $1 \le i \le n$, $X^{(i)}$ will denote the random vector $(X_i^{(i)}, \dots, X_n^{(i)})$. The marginal $P = G_1^n$ of $\overline{\Pi}$ will be the law of $X^{(1)} = (X_1^{(1)}, \dots, X_n^{(1)})$ and the marginal $Q = H_1^n$ of $\overline{\Pi}$ will be the law of $(Y_1, \dots, Y_n)$.

The structure of dependence between all these random variables is based on the following remark.

REMARK 1. Assume that X, Y, Z are three random variables. Assume that the law of (X, Y) admits the density $\sigma(x, y)$ with respect to $d\mu(x)\,d\nu(y)$, and that the law of (Y, Z) admits the density $\rho(y, z)$ with respect to $d\nu(y)\,d\lambda(z)$. Let $k(y)$ denote the density of the law of Y with respect to $d\nu(y)$. If the random variables X and Z are independent given Y, then the law of X, Y, Z admits the density

$$\frac{\sigma(x, y)\rho(y, z)}{k(y)},$$

with respect to $d\mu(x)\,d\nu(y)\,d\lambda(z)$.

Let us first consider the random variables $X_1^{(1)}, Y_1$. The law of $(X_1^{(1)}, Y_1)$ is given by its density l_1 that will be denoted

$$L_1\big(dx_1^{(1)}, dy_1\big) = l_1\big(x_1^{(1)}, y_1\big)\,\mu_1\big(dx_1^{(1)}\big)\,\mu_1(dy_1).$$

As in the case of dimension one, L_1 is defined so that the law of $X_1^{(1)}$ is G_1 and the law of Y_1 is H_1. Given $X_1^{(1)}$, Y_1, the law of

$$\left(X_2^{(1)}, \dots, X_n^{(1)}, X_2^{(2)}, \dots, X_n^{(2)}\right)$$

on $\mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ will be denoted

$$\begin{aligned}\Sigma_2^n&\left(dx_2^{(1)}, \dots, dx_n^{(1)}, dx_2^{(2)}, \dots, dx_n^{(2)} | X_1^{(1)}, Y_1\right)\\ &= \sigma_2^n\left(x_2^{(1)}, \dots, x_n^{(1)}, x_2^{(2)}, \dots, x_n^{(2)} | X_1^{(1)}, Y_1\right)\\ &\quad\times \mu_2\left(dx_2^{(1)}\right) \cdots \mu_n\left(dx_n^{(1)}\right) \mu_2\left(dx_2^{(2)}\right) \cdots \mu_n\left(dx_n^{(2)}\right).\end{aligned}$$

$\Sigma_2^n(\cdot\,|X_1^{(1)}, Y_1)$ is defined so that the law of $(X_2^{(1)}, \dots, X_n^{(1)})$ given $X_1^{(1)}, Y_1$ is $G_2^n(\cdot\,|X_1^{(1)})$, and the law of $(X_2^{(2)}, \dots, X_n^{(2)})$ given $X_1^{(1)}, Y_1$ is $G_2^n(\cdot\,|Y_1)$. To simplify the notation, for every $1 \le i \le n$ let $x^{(i)}$ denote the vector $(x_i^{(i)}, \dots, x_n^{(i)})$ on $\mathbb{R}^{n-i+1}$. The law of $(Y_1, X^{(1)}, X^{(2)})$ is given by the product density,

$$d_1\left(y_1, x^{(1)}, x^{(2)}\right) = l_1\left(x_1^{(1)}, y_1\right) \sigma_2^n\left(x_2^{(1)}, \dots, x_n^{(1)}, x_2^{(2)}, \dots, x_n^{(2)} | x_1^{(1)}, y_1\right).$$

In this construction, we easily see that the law of $X^{(1)}$ is $G_1^n = P$.

Now assume that for $2 \le i \le n$, the law of

$$\left(Y_1, \dots, Y_{i-1}, X^{(1)}, \dots, X^{(i)}\right)$$

is given by a density

$$d_{i-1}\left(y_1, \dots, y_{i-1}, x^{(1)}, \dots, x^{(i)}\right)$$

such that the law of $(Y_1, \dots, Y_{i-1})$, is H_1^{i-1} and the law of the random vector $X^{(i)}$ given $Y_1, \dots, Y_{i-1}$ is

$$G_i^n(\cdot\,|Y_1, \dots, Y_{i-1}).$$

Then we first introduce the random variable Y_i for $2 \le i \le n$. The law of $(X_i^{(i)}, Y_i)$ given $Y_1, \dots, Y_{i-1}$ will be denoted

$$L_i(dx_i^{(i)}, dy_i | Y_1, \dots, Y_{i-1}) = l_i(x_i^{(i)}, y_i | Y_1, \dots, Y_{i-1})\, \mu_i(dx_i^{(i)})\, \mu_i(dy_i).$$

$L_i(\cdot\,|Y_1, \dots, Y_{i-1})$ is defined so that the law of Y_i given $Y_1, \dots, Y_{i-1}$ is

$$H_i(\cdot\,|Y_1, \dots, Y_{i-1}),$$

and the law of $X_i^{(i)}$ given $Y_1, \dots, Y_{i-1}$ is

$$G_i(\cdot\,|Y_1, \dots, Y_{i-1}).$$

Using Remark 1, the density of the law of

$$(Y_1, \dots, Y_{i-1}, Y_i, X^{(1)}, \dots, X^{(i)})$$

will be given by the density

$$d_i(y_1, \dots, y_i, x^{(1)}, \dots, x^{(i)}) \tag{2.35}$$
$$= \frac{d_{i-1}(y_1, \dots, y_{i-1}, x^{(1)}, \dots, x^{(i)})\; l_i(x_i^{(i)}, y_i | y_1, \dots, y_{i-1})}{g_i(x_i^{(i)} | y_1, \dots, y_{i-1})}.$$

Thus, according to Remark 1, Y_i is independent of

$$X^{(1)}, \dots, X^{(i-1)}, (X_{i+1}^{(i)}, \dots, X_n^{(i)})$$

given $X_i^{(i)}$, $Y_1, \dots, Y_{i-1}$. For $1 \le i \le n-1$, we then introduce the random vector $X^{(i+1)}$. The law of

$$\big((X_{i+1}^{(i)}, \dots, X_n^{(i)}), (X_{i+1}^{(i+1)}, \dots, X_n^{(i+1)})\big)$$

on $\mathbb{R}^{n-i} \times \mathbb{R}^{n-i}$ given $Y_1, \dots, Y_i$, $X_i^{(i)}$ will be denoted

$$\begin{aligned}\Sigma_{i+1}^n(dx_{i+1}^{(i)}, \dots, dx_n^{(i)}, dx_{i+1}^{(i+1)}, \dots, dx_n^{(i+1)} | X_i^{(i)}, Y_1, \dots, Y_i)\\ = \sigma_{i+1}^n(x_{i+1}^{(i)}, \dots, x_n^{(i)}, x_{i+1}^{(i+1)}, \dots, x_n^{(i+1)} | X_i^{(i)}, Y_1, \dots, Y_i)\\ \times \mu_{i+1}(dx_{i+1}^{(i)}) \cdots \mu_n(dx_n^{(i)})\, \mu_{i+1}(dx_{i+1}^{(i+1)}) \cdots \mu_n(dx_n^{(i+1)}).\end{aligned}$$

$\Sigma_{i+1}^n(\cdot\, | X_i^{(i)}, Y_1, \dots, Y_i)$ will be defined so that the law of

$$\big(X_{i+1}^{(i)}, \dots, X_n^{(i)}\big)$$

conditionally on $Y_1, \dots, Y_i$, $X_i^{(i)}$ is

$$G_{i+1}^n\big(\cdot\, | Y_1, \dots, Y_{i-1}, X_i^{(i)}\big),$$

and the law of $X^{(i+1)}$ conditionally on $Y_1, \dots, Y_i$, $X_i^{(i)}$ is

$$G_{i+1}^n(\cdot\, | Y_1, \dots, Y_{i-1}, Y_i).$$

Using Remark 1, the density of the law of

$$(Y_1, \dots, Y_i, X^{(1)}, \dots, X^{(i)}, X^{(i+1)})$$

is given by the density

$$\begin{aligned}&d_i(y_1, \dots, y_i, x^{(1)}, \dots, x^{(i)}, x^{(i+1)})\\ &\quad = \frac{\bar{d}_i(y_1, \dots, y_i, x^{(1)}, \dots, x^{(i)})\; \sigma_{i+1}^n(x_{i+1}^{(i)}, \dots, x_n^{(i)}, x_{i+1}^{(i+1)}, \dots, x_n^{(i+1)} | x_i^{(i)}, y_1^i)}{g_{i+1}^n(x_{i+1}^{(i)}, \dots, x_n^{(i)} | y_1, \dots, y_{i-1}, x_i^{(i)})}.\end{aligned}$$

Thus, according to Remark 1, $X^{(i+1)}$ is independent of $X^{(1)}, \dots, X^{(i-1)}$ given $X^{(i)}$, $Y_1, \dots, Y_i$.

In this way, by induction over i, we construct the law $\overline{\Pi}$ of the family of random variables,

$$Y_1, \dots, Y_n, X^{(1)}, \dots, X^{(n)},$$

so that the law $\overline{\Pi}$ is given by the density

$$\overline{\pi} = \bar{d}_n(y_1, \dots, y_n, x^{(1)}, \dots, x^{(n)}).$$

Now, we will set with more details the expression of $\overline{\pi}$. Let us first give the exact expression of the density $l_i(\cdot|y_1^{i-1})$, $1 \le i \le n$. For every $1 \le i \le n$, we have

$$\begin{aligned} l_i(x_i^{(i)}, y_i|y_1^{i-1}) &\\ = \mathbf{1}_{x_i^{(i)}=y_i} &\min(h_i(y_i|y_1^{i-1}), g_i(x_i^{(i)}|y_1^{i-1})) \\ &+ \mathbf{1}_{x_i^{(i)}\neq y_1} \frac{[h_i(y_i|y_1^{i-1}) - g_i(y_i|y_1^{i-1})]_+ [g_i(x_i|y_1^{i-1}) - h_i(x_i|y_1^{i-1})]_+}{\|H_i(\cdot|y_1^{i-1}) - G_i(\cdot|y_i^{i-1})\|_{\mathrm{TV}}}. \end{aligned}$$

As for the case $n = 1$, integrating one of the variables $x_i^{(i)}$ or y_i, it is clear that the marginals of $L_i(\cdot|y_1^{i-1})$ are $G_i(\cdot|y_1^{i-1})$ and $H_i(\cdot|y_1^{i-1})$.

Now, let us describe, for every $2 \le i \le n$, the measure

$$\Sigma_i^n\big(\cdot\,|x_{i-1}^{(i-1)}, y_1, \dots, y_{i-1}\big).$$

To present the condition satisfied by $\Sigma_i^n(\cdot\,|x_{i-1}^{(i-1)}, y_1, \dots, y_{i-1})$, we need the following result by Fiebig [6] [see inequality (2.1), page 482].

PROPOSITION 2. *Let Q and R be two probability measures on $\mathbb{R}^k$ with strictly positive densities q and r with respect to a measure ν on $\mathbb{R}^k$. Let $(Z_1, \dots, Z_k)$ (resp. $(W_1, \dots, W_k)$) be a random vector on $\mathbb{R}^k$ whose law is Q (resp. R). Then, there exists a probability measure whose density is σ with respect of $\nu \otimes \nu$ on $\mathbb{R}^k \times \mathbb{R}^k$ such that, for every $1 \le j \le k$,*

$$\iint \mathbf{1}_{z_j \neq w_j}\, \sigma(z, w)\, d\nu(z)\, d\nu(w) \le \|Q - R\|_{\mathrm{TV}}.$$

Fiebig proves this result for probability measures on a countable set S. The proof is easily extended to probability measures on $\mathbb{R}^k$ with strictly positive densities yielding thus Proposition 2. Thanks to Proposition 2, we may assume that, for every $2 \le i \le n$, $\Sigma_i^n(\cdot, \dots, \cdot|x_{i-1}^{(i-1)}, y_1^{i-1})$ satisfies the following conditions. For every $2 \le i \le n$, the marginals of

$$\Sigma_i^n\big(dx_i^{(i-1)}, \dots, dx_n^{(i-1)}, dx_i^{(i)}, \dots, dx_n^{(i)}|x_{i-1}^{(i-1)}, y_1, \dots, y_{i-1}\big)$$

are

$$G_i^n\big(dx_i^{(i-1)}, \dots, dx_n^{(i-1)}|y_1, \dots, y_{i-2}, x_{i-1}^{(i-1)}\big)$$

and

$$G_i^n\big(dx_i^{(i)}, \dots, dx_n^{(i)}|y_1, \dots, y_{i-2}, y_{i-1}\big).$$

Recall that if $X = (X_1, \dots, X_n)$ is a sample whose law is P,

$$G_i^n\big(\cdot\,|y_1, \dots, y_{i-2}, x_{i-1}^{(i-1)}\big)$$

is the law of $(X_i, \dots, X_n)$ given $X_{i-1} = x_{i-1}^{(i-1)}$ and $X_1^{i-2} = y_1^{i-2}$. Similarly,

$$G_i^n(\cdot\,|y_1, \dots, y_{i-2}, y_{i-1})$$

is the law of $(X_i, \dots, X_n)$ given $X_{i-1} = y_{i-1}$ and $X_1^{i-2} = y_1^{i-2}$. According to Proposition 2,

$$\Sigma_i^n(\cdot \dots, \cdot|x_{i-1}^{(i-1)}, y_1^{i-1})$$

satisfies the following additional property, for every $2 \le i \le j \le n$,

$$\begin{aligned} &\int \cdots \int \mathbf{1}_{x_j^{(i-1)} \neq x_j^{(i)}} \Sigma_i^n\big(dx_i^{(i-1)}, \dots, dx_n^{(i-1)}, dx_i^{(i)}, \dots, dx_n^{(i)}|x_{i-1}^{(i-1)}, y_1^{i-1}\big) \\ &\qquad \le a_j\big(y_1^{i-2}, x_{i-1}^{(i-1)}, y_{i-1}\big), \end{aligned} \tag{2.36}$$

where

$$\begin{aligned} &a_j\big(y_1^{i-2}, x_{i-1}^{(i-1)}, y_{i-1}\big) \\ &\qquad = \|\mathscr{L}\big(X_j^n|X_1^{i-1} = y_1^{i-1}\big) - \mathscr{L}\big(X_j^n|X_1^{i-2} = y_1^{i-2}, X_{i-1} = x_{i-1}^{(i-1)}\big)\|_{\mathrm{TV}}. \end{aligned}$$

Let us now present the expression of $\overline{\pi}$. For $2 \le i \le n$, define

$$\begin{aligned} &\xi_i^n\big(y_i, x_i^{(i-1)}, \dots, x_n^{(i-1)}, x_i^{(i)}, \dots, x_n^{(i)}|x_{i-1}^{(i-1)}, y_1^{i-1}\big) \\ &\qquad = \frac{\sigma_i^n(x_i^{(i-1)}, \dots, x_n^{(i-1)}, x_i^{(i)}, \dots, x_n^{(i)}|x_{i-1}^{(i-1)}, y_1^{i-1})}{g_i^n(x_i^{(i)}, \dots, x_n^{(i)}|y_1^{i-1})} l_i(x_i^{(i)}, y_i|y_1^{i-1}). \end{aligned}$$

We have

$$\begin{aligned} &\overline{\pi}\big(y_1, \dots, y_n, x^{(1)}, \dots, x^{(n)}\big) \\ &\qquad = \overline{d}_n\big(y_1, \dots, y_n, x^{(1)}, \dots, x^{(n)}\big) \\ &\qquad = l_1(x_1^{(1)}, y_1) \prod_{i=2}^n \xi_i^n\big(y_i, x_i^{(i-1)}, \dots, x_n^{(i-1)}, x_i^{(i)}, \dots, x_n^{(i)}|x_{i-1}^{(i-1)}, y_1^{i-1}\big). \end{aligned}$$

This density $\overline{\pi}$ has all the properties to be a good candidate to prove (2.28) and (2.29). Indeed, integrating successively $\overline{\pi}$ with respect to the variables

$$(x_2^{(1)}, \dots, x_n^{(1)}), (x_3^{(2)}, \dots, x_n^{(2)}), \dots, x_{n-1}^{(n)},$$

and then with respect to the variables

$$x_n^{(n)}, \dots, x_1^{(1)},$$

we see that the law of $(Y_1, \dots, Y_n)$ is Q. Similarly, integrating successively with respect to

$$y_n, x^{(n)}, y_{n-1}, x^{(n-1)}, \dots, y_2, x^{(2)}, y_1,$$

shows that the law of the random vector $X^{(1)}$ is P. Therefore, our aim is to prove that for every $1 \leq j \leq n$,

$$\int \cdots \int \alpha_j(y_1, \dots, y_n) \mathbf{1}_{y_j \neq x_j^{(1)}} \, d\overline{\Pi}(y_1, \dots, y_n, x^{(1)}, \dots, x^{(n)}) \leq \sum_{i=1}^{j} \gamma_i^j (2E_i)^{1/2} (\Delta_j)^{1/2} \tag{2.37}$$

and

$$\int \cdots \int \beta_j(x_1^{(1)}, \dots, x_n^{(1)}) \mathbf{1}_{y_j \neq x_j^{(1)}} \, d\overline{\Pi}(y_1, \dots, y_n, x^{(1)}, \dots, x^{(n)}) \leq \sum_{i=1}^{j} \gamma_i^j (2E_i)^{1/2} (\widetilde{\Delta}_j)^{1/2}. \tag{2.38}$$

Equations (2.37) and (2.38) are very similar and the scheme of their proof is identical. First, we present the proof of (2.37) and then of (2.38). For every $1 \leq j \leq n$, we have,

$$\mathbf{1}_{y_j \neq x_j^{(1)}} \leq \mathbf{1}_{y_j \neq x_j^{(j)}} + \mathbf{1}_{x_j^{(j)} \neq x_j^{(j-1)}} + \cdots + \mathbf{1}_{x_j^{(2)} \neq x_j^{(1)}}.$$

Hence, we get

$$\int \cdots \int \alpha_j(y_1, \dots, y_n) \mathbf{1}_{y_j \neq x_j^{(1)}} \, d\overline{\Pi}(y_1, \dots, x^{(n)}) \leq A_j + \sum_{i=1}^{j-1} B_j^{(i)},$$

where

$$A_j = \int \cdots \int \alpha_j(y_1, \dots, y_n) \mathbf{1}_{y_j \neq x_j^{(j)}} \, d\overline{\Pi}(y_1, \dots, x^{(n)})$$

and

$$B_j^{(i)} = \int \cdots \int \alpha_j(y_1, \dots, y_n) \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i)}} \, d\overline{\Pi}(y_1, \dots, x^{(n)}).$$

Thus, the proof of (2.37) is now divided in two parts, the study of the integral A_j and then the study of the integral $B_j^{(i)}$. Integrating successively the density $\overline{\pi}$ with respect to the variables

$$(x_2^{(1)}, \dots, x_n^{(1)}), (x_3^{(2)}, \dots, x_n^{(2)}), \dots, x_{n-1}^{(n)},$$

we show that the law of

$$(Y_1, \dots, Y_n, X_1^{(1)}, \dots, X_n^{(n)})$$

is given by the density

$$l_1(x_1^{(1)}, y_1) \cdots l_n(x_n^{(n)}, y_n | y_1^{n-1}).$$

Consequently, the law of $(Y_1, \dots, Y_n, X_j^{(j)})$ is

$$H_{j+1}^n(dy_{j+1}, \dots, dy_n | y_1^j) \, L_j(dx_j^{(j)}, dy_j | y_1^{j-1}) \, H_1^{j-1}(dy_1, \dots, dy_{j-1}).$$

Thus, for every $1 \le j \le n$,

$$A_j = \iint \Big(\int \alpha_j(y_1, \dots, y_n)\, H_{j+1}^n(dy_{j+1}, \dots, dy_n | y_1^j) \Big) \times \mathbf{1}_{y_j \neq x_j^{(j)}} L_j(dx_j^{(j)}, dy_j | y_1^{j-1}) H_1^{j-1}(dy_1, \dots, dy_{j-1}).$$

From the definition of l_j, we have

$$\mathbf{1}_{y_j \neq x_j^{(j)}} l_j(x_j^{(j)}, y_j | y_1^{j-1}) = \mathbf{1}_{y_j \neq x_j^{(j)}} \frac{[h_j(y_j|y_1^{j-1}) - g_j(y_j|y_1^{j-1})]_+ [g_j(x_j^{(j)}|y_1^{j-1}) - h_j(x_j^{(j)}|y_1^{j-1})]_+}{\|H_j(\cdot\,|y_1^{j-1}) - G_j(\cdot|y_1^{j-1})\|_{\mathrm{TV}}}.$$

Integrating with respect to $x_j^{(j)}$, it follows that

$$A_j = \iint \Big(\int \alpha_j(y_1, \dots, y_n)\, H_{j+1}^n(dy_{j+1}, \dots, dy_n | y_1^j) \Big) \times \big[h_j(y_j|y_1^{j-1}) - g_j(y_j|y_1^{j-1})\big]_+ \, \mu_j(dy_j)\, H_1^{j-1}(dy_1, \dots, dy_{j-1}).$$

Then, by the Cauchy–Schwarz inequality,

$$A_j \le \int \Big(\int \alpha_j(y_1, \dots, y_n)^2 \, H_j^n(dy_j, \dots, dy_n | y_1^{j-1}) \Big)^{1/2} \times \left(\int \left[1 - \frac{g_j(y_j|y_1^{j-1})}{h_j(y_j|y_1^{j-1})} \right]_+^2 H_j(dy_j|y_1^{j-1}) \right)^{1/2} H_1^{j-1}(dy_1, \dots, dy_{j-1}).$$

According to its definition,

$$d_{\mu_j}(g_j(\cdot|y_1^{j-1}) | h_j(\cdot\,|y_1^{j-1})) = \left(\int \left[1 - \frac{g_j(y_j|y_1^{j-1})}{h_j(y_j|y_1^{j-1})} \right]_+^2 H_j(dy_j|y_1^{j-1}) \right)^{1/2}.$$

From the inequality (2.32) of Lemma 2, we have

$$d_{\mu_j}(g_j(\cdot|y_1^{j-1}) | h_j(\cdot\,|y_1^{j-1})) \le \left(2\,\mathrm{Ent}_{G_j(\cdot\,|y_1^{j-1})} \left(\frac{h_j(\cdot\,|y_1^{j-1})}{g_j(\cdot\,|y_1^{j-1})} \right) \right)^{1/2}.$$

By the Cauchy–Schwarz inequality again, it follows that

$$A_j \le \Big(\int \alpha_j(y_1, \dots, y_n)^2 \, Q(dy_1, \dots, dy_n) \Big)^{1/2} \times \left(\int 2\,\mathrm{Ent}_{G_j(\cdot\,|y_1^{j-1})} \left(\frac{h_j(\cdot\,|y_1^{j-1})}{g_j(\cdot\,|y_1^{j-1})} \right) H_1^{j-1}(dy_1, \dots, dy_{j-1}) \right)^{1/2}.$$

Finally, with the definitions of E_j and Δ_j, we get

$$A_j \le (2E_j)^{1/2} (\Delta_j)^{1/2}. \tag{2.39}$$

We now want to bound similarly $B_j^{(j)}$. Recall that

$$B_j^{(i)} = \int \cdots \int \alpha_j(y_1, \dots, y_n) \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i)}} \overline{\Pi}(dy_1, \dots, dx^{(n)}).$$

Let here $\hat{\Pi}(\cdot\,|x^{(i)}, x^{(i+1)}, y_1^i)$ denote the law of $(Y_{i+1}, \dots, Y_n)$ given

$$X^{(i)} = x^{(i)}, \qquad X^{(i+1)} = x^{(i+1)}, \qquad Y_1^i = y_1^i.$$

Integrating successively the density $\overline{\pi}$ with respect to the variables

$$\big(x_2^{(1)}, \dots, x_n^{(1)}\big), \dots, \big(x_i^{(i-1)}, \dots, x_n^{(i-1)}\big), x_1^{(1)}, \dots, x_{i-1}^{(i-1)},$$

and then with respect to

$$y_n, x^{(n)}, y_{n-1}, x^{(n-1)}, \dots, y_{i+2}, x^{(i+2)},$$

we see that the law of $(X^{(i)}, X^{(i+1)}, Y_1, \dots, Y_i)$ is given by

$$\begin{aligned} &\Sigma_{i+1}^n(dx_{i+1}^{(i)}, \dots, dx_n^{(i)}, dx_{i+1}^{(i+1)}, \dots, dx_n^{(i+1)}|x_i^{(i)}, y_1^i) \\ &\quad \times L_i(dx_i^{(i)}, dy_i|y_1^{i-1})\, H_1^{i-1}(dy_1, \dots, dy_{i-1}). \end{aligned}$$

Thus, we have

$$\begin{aligned} B_j^{(i)} = \iint &\alpha_j(y_1, \dots, y_n) \hat{\Pi}(dy_{i+1}, \dots, dy_n|x^{(i)}, x^{(i+1)}, y_1^i) \\ &\times \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i)}} \Sigma_{i+1}^n(dx_{i+1}^{(i)}, \dots, dx_n^{(i)}, dx_{i+1}^{(i+1)}, \dots, dx_n^{(i+1)}|x_i^{(i)}, y_1^i) \\ &\times L_i(dx_i^{(i)}, dy_i|y_1^{i-1})\, H_1^{i-1}(dy_1, \dots, dy_{i-1}). \end{aligned}$$

Consequently, by Cauchy–Schwarz inequality,

$$\begin{aligned} B_j^{(i)} \le \int &\Big(\int \alpha_j(y_1, \dots, y_n)^2 \widetilde{\Pi}(dy_{i+1}, \dots, dy_n|x_i^{(i)}, y_1^i)\Big)^{1/2} \\ &\times \left(\int \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i)}} \Sigma_{i+1}^n(dx_{i+1}^{(i)}, \dots, dx_n^{(i)}, dx_{i+1}^{(i+1)}, \dots, dx_n^{(i+1)}|x_i^{(i)}, y_1^i)\right)^{1/2} \\ &\times L_i(dx_i^{(i)}, dy_i|y_1^{i-1})\, H_1^{i-1}(dy_1, \dots, dy_{i-1}), \end{aligned}$$

where $\widetilde{\Pi}(\cdot|x_i^{(i)}, y_1^i)$ denotes the law of $(Y_{i+1}, \dots, Y_n)$ given

$$X_i^{(i)} = x_i^{(i)}, \qquad Y_1^i = y_1^i.$$

Actually, we have

$$\widetilde{\Pi}(\cdot\,|x_i^{(i)}, y_1^i) = H_{i+1}^n(\cdot\,|y_1^i)$$

and therefore $\widetilde{\Pi}$ is independent of $x_i^{(i)}$. By the property (2.36) of the measure Σ_{i+1}^n, we know that

$$\begin{aligned} &\int \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i)}} \Sigma_{i+1}^n(dx_{i+1}^{(i)}, \dots, dx_n^{(i)}, dx_{i+1}^{(i+1)}, \dots, dx_n^{(i+1)}|x_i^{(i)}, y_1^i) \\ &\qquad \le a_j(y_1^{i-1}, x_i^{(i)}, y_i). \end{aligned}$$

Thanks to this inequality, we bound $B_j^{(i)}$, either with the coefficient γ_i^j, or with $\widetilde{\gamma}_i^j$, as follows. By definition, we know that for every real number $y_1, \dots, y_i, x_i^{(i)}$,

$$a_j(y_1^{i-1}, x_i^{(i)}, y_i) \le (\gamma_i^j)^2 \mathbf{1}_{x_i^{(i)} \neq y_i}. \tag{2.40}$$

Therefore

$$\begin{aligned} B_j^{(i)} \le \gamma_i^j \int \Big(\int \alpha_j(y_1, \dots, y_n)^2 H_{i+1}^n(dy_{i+1}, \dots, dy_n | y_1^i) \Big)^{1/2} \\ \times \mathbf{1}_{x_i^{(i)} \neq y_i} L_i(dx_i^{(i)}, dy_i | y_1^{i-1}) H_1^{i-1}(dy_1, \dots, dy_{i-1}). \end{aligned} \tag{2.41}$$

The second way to bound $B_j^{(i)}$ with $\widetilde{\gamma}_i^j$ is quite different since we do not take the supremum over all $y_1, \dots, y_i, x_i^{(i)}$ in $\mathbb{R}$. By the triangular inequality applied to the norm $\|\cdot\|_{\mathrm{TV}}$, we have

$$a_j(y_1^{i-1}, x_i^{(i)}, y_i) \le (\tilde{a}_j(y_1^{i-1}, x_i^{(i)}) + \tilde{a}_j(y_1^{i-1}, y_i)) \mathbf{1}_{x_i^{(i)} \neq y_i},$$

where

$$\tilde{a}_j(y_1^{i-1}, y_i) = \| \mathscr{L}(X_j^n | X_1^i = y_1^i) - \mathscr{L}(X_j^n) \|_{\mathrm{TV}}.$$

The density $g_i(\cdot | y_1^{i-1})$ is strictly positive. Therefore, the measure

$$L_i(dx_i^{(i)}, dy_i | y_1^{i-1})$$

is absolutely continuous with respect to the measure

$$G_i(dx_i^{(i)} | y_1^{i-1}) \, G_i(dy_i | y_1^{i-1}).$$

Moreover, the measure

$$H_1^{i-1}(dy_1, \dots, dy_{i-1})$$

is absolutely continuous with respect to the measure

$$G_1^{i-1}(dy_1, \dots, dy_{i-1}),$$

since g_1^{i-1} is a strictly positive density. It follows that the measure

$$L_i(dx_i^{(i)}, dy_i | y_1^{i-1}) \, H_1^{i-1}(dy_1, \dots, dy_{i-1})$$

is absolutely continuous with respect to the measure

$$G_i(dx_i^{(i)} | y_1^{i-1}) \, G_i(dy_i | y_1^{i-1}) \, G_1^{i-1}(dy_1, \dots, dy_{i-1}).$$

According to the definition of $\widetilde{\gamma}_i^j$, it follows that

$$\tilde{a}_j(y_1^{i-1}, y_i) \le \tfrac{1}{2} (\widetilde{\gamma}_i^j)^2$$

for almost every y_1^i with respect to the measure G_1^i, the law of X_1^i. Therefore, for almost every y_1^{i-1}, y_i, $x_i^{(i)}$ with respect to the measure

$$G_i(dx_i^{(i)}|y_1^{i-1})\,G_i(dy_i|y_1^{i-1})\,G_1^{i-1}(dy_1,\dots,dy_{i-1}),$$

we have

$$a_j(y_1^{i-1},x_i^{(i)},y_i)\le(\tilde{\gamma}_i^j)^2\mathbf{1}_{x_i^{(i)}\neq y_i}. \tag{2.42}$$

This inequality is still true for almost every y_1^{i-1}, y_i, $x_i^{(i)}$ with respect to the measure.

$$L_i(dx_i^{(i)},dy_i|y_1^{i-1})H_1^{i-1}(dy_1,\dots,dy_{i-1}).$$

It follows that

$$\begin{aligned}B_j^{(i)}\le\tilde{\gamma}_i^j\int\Bigl(\int\alpha_j(y_1,\dots,y_n)^2\,H_{i+1}^n(dy_{i+1},\dots,dy_n|y_1^i)\Bigr)^{1/2}\\ \times\mathbf{1}_{x_i^{(i)}\neq y_i}L_i(dx_i^{(i)},dy_i|y_1^{i-1})\,H_1^{i-1}(dy_1,\dots,dy_{i-1}).\end{aligned} \tag{2.43}$$

The end of the proof is obviously the same from inequality (2.41) or (2.43). From (2.41), integrating with respect to the variable $x_i^{(i)}$, we obtain

$$\begin{aligned}B_j^{(i)}\le\gamma_i^j\int\Bigl(\int\alpha_j(y_1,\dots,y_n)^2\,H_{i+1}^n(dy_{i+1},\dots,dy_n|y_1^i)\Bigr)^{1/2}\\ \times\bigl[h_i(y_i|y_1^{i-1})-g_i(y_i|y_1^{i-1})\bigr]_+\,\mu_i(dy_i)\,H_1^{i-1}(dy_1,\dots,dy_{i-1}).\end{aligned}$$

Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned}B_j^{(i)}\le\gamma_i^j\int\Bigl(\int\alpha_j(y_1,\dots,y_n)^2\,H_i^n(dy_i,\dots,dy_n|y_1^{i-1})\Bigr)^{1/2}\\ \times\biggl(\int\biggl[1-\frac{g_i(y_i|y_1^{i-1})}{h_i(y_i|y_1^{i-1})}\biggr]_+^2H_i(dy_i|y_1^{i-1})\biggr)^{1/2}H_1^{i-1}(dy_1,\dots,dy_{i-1}).\end{aligned}$$

We finish as for the bound of A_j, applying (2.32) of Lemma 2. We thus get

$$B_j^{(i)}\le\gamma_i^j(\Delta_j)^{1/2}(2E_i)^{1/2}. \tag{2.44}$$

From (2.44) and (2.39), we deduce (2.37). This ends of proof of (2.11) of Theorem 1.

As we already mentioned, the scheme of the proof of inequality (2.38) is the same as the one of inequality (2.37). For every $1\le j\le n$,

$$\mathbf{1}_{y_j\neq x_j^{(1)}}\le\mathbf{1}_{y_j\neq x_j^{(j)}}+\mathbf{1}_{x_j^{(j)}\neq x_j^{(j-1)}}+\cdots+\mathbf{1}_{x_j^{(2)}\neq x_j^{(1)}}.$$

Hence,

$$\int\cdots\int\beta_j(x_1^{(1)},\dots,x_n^{(1)})\mathbf{1}_{y_j\neq x_j^{(1)}}\overline{\Pi}(dy_1,\dots,dx^{(n)})\le C_j+\sum_{i=1}^{j-1}D_j^{(i)},$$

where

$$C_j = \int \cdots \int \beta_j(x_1^{(1)}, \dots, x_n^{(1)}) \mathbf{1}_{y_j \neq x_j^{(j)}} \overline{\Pi}(dy_1, \dots, dx^{(n)})$$

and

$$D_j^{(i)} = \int \cdots \int \beta_j(x_1^{(1)}, \dots, x_n^{(1)}) \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i+1)}} \overline{\Pi}(dy_1, \dots, dx^{(n)}).$$

First we study the integral C_j and then the integral $D_j^{(i)}$. Let $\breve{\Pi}_j(\cdot\,|x_j^{(j)}, y_1^j)$ denote the law of

$$(X^{(1)}, \dots, X^{(j-1)}, (X_{j+1}^{(j)}, \dots, X_n^{(j)}))$$

given

$$X_j^{(j)} = x_j^{(j)}, \qquad Y_1^j = y_1^j.$$

This law is independent of y_j. Indeed, as we already deduced from (2.35),

$$(X^{(1)}, \dots, X^{(j-1)}, (X_{j+1}^{(j)}, \dots, (X_n^{(j)}))$$

is independent of Y_j given $X_j^{(j)}, Y_1, \dots, Y_{j-1}$. We therefore denote

$$\breve{\Pi}_j(\cdot\,|x_j^{(j)}, y_1, \dots, y_j) = \breve{\Pi}_j(\cdot\,|x_j^{(j)}, y_1^{j-1}).$$

The law of $(X_j^{(j)}, Y_1, \dots, Y_j)$ is given by

$$L_j(dx_j^{(j)}, dy_j|y_1^{j-1}) H_1^{j-1}(dy_1, \dots, dy_{j-1}).$$

Therefore,

$$\begin{aligned} C_j = \iint \beta_j(x_1^{(1)}, \dots, x_n^{(1)})\, \breve{\Pi}_j(dx_1^{(1)}, \dots, dx_n^{(j)}|x_j^{(j)}, y_1^{j-1}) \\ \times \mathbf{1}_{y_j \neq x_j^{(j)}} L_j(dx_j^{(j)}, dy_j|y_1^{j-1})\, H_1^{j-1}(dy_1, \dots, dy_{j-1}). \end{aligned}$$

From the definition of l_j, we have

$$\begin{aligned} &\mathbf{1}_{y_j \neq x_j^{(j)}} l_j(x_j^{(j)}, y_j|y_1^{j-1}) \\ &\quad = \mathbf{1}_{y_j \neq x_j^{(j)}} \frac{[h_j(y_j|y_1^{j-1}) - g_j(y_j|y_1^{j-1})]_+ [g_j(x_j^{(j)}|y_1^{j-1}) - h_j(x_j^{(j)}|y_1^{j-1})]_+}{\|H_j(\cdot\,|y_1^{j-1}) - G_j(\cdot|y_1^{j-1})\|_{\mathrm{TV}}}. \end{aligned}$$

Integrating with respect to y_j, it follows that

$$\begin{aligned} C_j = \iint \Big(\int \beta_j(x_1^{(1)}, \dots, x_n^{(1)})\, \breve{\Pi}_j(dx_1^{(1)}, \dots, dx_n^{(j)}|x_j^{(j)}, y_1^{j-1}) \Big) \\ \times [g_j(x_j^{(j)}|y_1^{j-1}) - h_j(x_j^{(j)}|y_1^{j-1})]_+\, \mu_j(dx_j^{(j)})\, H_1^{j-1}(dy_1, \dots, dy_{j-1}). \end{aligned}$$

Then, by the Cauchy–Schwarz inequality,

$$C_j \le \int \Big(\int \beta_j\big(x_1^{(1)},\dots,x_n^{(1)}\big)^2\, \widetilde{\Pi}_j\big(dx_1^{(1)},\dots,dx_n^{(j)}|x_j^{(j)},y_1^{j-1}\big)\, G_j\big(dx_j^{(j)}|y_1^{j-1}\big)\Big)^{1/2}$$
$$\times \left(\int\left[1-\frac{h_j(x_j^{(j)}|y_1^{j-1})}{g_j(x_j^{(j)}|y_1^{j-1})}\right]_+^2 G_j\big(dx_j^{(j)}|y_1^{j-1}\big)\right)^{1/2} H_1^{j-1}(dy_1,\dots,dy_{j-1}).$$

By definition,

$$d_{\mu_j}\big(h_j(\cdot|y_1^{j-1})|g_j(\cdot\,|y_1^{j-1})\big) = \left(\int\left[1-\frac{h_j(x_j^{(j)}|y_1^{j-1})}{g_j(x_j^{(j)}|y_1^{j-1})}\right]_+^2 G_j(dx_j^{(j)}|y_1^{j-1})\right)^{1/2}.$$

From the inequality (2.32) of Lemma 2, we have that

$$d_{\mu_j}\big(h_j(\cdot|y_1^{j-1})|g_j(\cdot\,|y_1^{j-1})\big) \le \left(2\,\mathrm{Ent}_{G_j(\cdot\,|y_1^{j-1})}\left(\frac{h_j(\cdot\,|y_1^{j-1})}{g_j(\cdot\,|y_1^{j-1})}\right)\right)^{1/2}.$$

By the Cauchy–Schwarz inequality, it follows that

$$C_j \le \Big(\int \beta_j(x_1^{(1)},\dots,x_n^{(1)})^2\, P(dx_1^{(1)},\dots,dx_n^{(1)})\Big)^{1/2}$$
$$\times \left(\int 2\,\mathrm{Ent}_{G_j(\cdot\,|y_1^{j-1})}\left(\frac{h_j(\cdot\,|y_1^{j-1})}{g_j(\cdot\,|y_1^{j-1})}\right) H_1^{j-1}(dy_1,\dots,dy_{j-1})\right)^{1/2}.$$

Finally, from the definition of E_j and $\widetilde{\Delta}_j$, we get

$$C_j \le (2E_j)^{1/2}(\widetilde{\Delta}_j)^{1/2}. \tag{2.45}$$

Now we will bound $D_j^{(i)}$ with the same tools as for the bound of $B_j^{(i)}$. Let

$$\breve{\Pi}_i(\cdot\,|x^{(i)},x^{(i+1)},y_1^i)$$

denote the law of $(X^{(1)},\dots,X^{(i-1)})$, given

$$\big(X^{(i)}=x^{(i)},\ X^{(i+1)}=x^{(i+1)},\ Y_1^i=y_1^i\big).$$

The law of $(X^{(i)},X^{(i+1)},Y_1^i)$ is

$$\Sigma_{i+1}^n\big(dx_{i+1}^{(i)},\dots,dx_n^{(i)},dx_{i+1}^{(i+1)},\dots,dx_n^{(i+1)}|x_i^{(i)},y_1^i\big)$$
$$\times L_i(dx_i^{(i)},dy_i|y_1^{i-1})\, H_1^{i-1}(dy_1,\dots,dy_{i-1}).$$

Let

$$T_j(x_i^{(i)},y_1^i)$$
$$= \int \mathbf{1}_{x_j^{(i+1)}\neq x_j^{(i)}}\Sigma_{i+1}^n\big(dx_{i+1}^{(i)},\dots,dx_n^{(i)},dx_{i+1}^{(i+1)},\dots,dx_n^{(i+1)}|x_i^{(i)},y_1^i\big)$$

and

$$U_j(x_i^{(i)}, y_1^i) = \int \beta_j(x_1^{(1)}, \dots, x_n^{(1)})^2 \check{\Pi}_i(dx_1^{(1)}, \dots, dx_n^{(i-1)} | x^{(i)}, x^{(i+1)}, y_1^i)$$

$$\times \Sigma_{i+1}^n(dx_{i+1}^{(i)}, \dots, dx_n^{(i)}, dx_{i+1}^{(i+1)}, \dots, dx_n^{(i+1)} | x_i^{(i)}, y_1^i).$$

Recall the definition of $D_j^{(i)}$,

$$D_j^{(i)} = \int \cdots \int \beta_j(x_1^{(1)}, \dots, x_n^{(1)}) \mathbf{1}_{x_j^{(i+1)} \neq x_j^{(i+1)}} \overline{\Pi}(dy_1, \dots, dx^{(n)}).$$

By the Cauchy–Schwarz inequality, we have

$$D_j^{(i)} \le \iint \big(U_j(x_i^{(i)}, y_1^i)\big)^{1/2} \big(T_j(x_i^{(i)}, y_1^i)\big)^{1/2}$$

$$\times L_1(dx_i^{(i)}, dy_i | y_1^{i-1}) \, H_1^{i-1}(dy_1, \dots, dy_{i-1}).$$

Actually, $U_j(x_i^{(i)}, y_1^i)$ is independent of y_i. Indeed, integrating with respect to the variables $x_{i+1}^{(i+1)}, \dots, x_n^{(i+1)}$, we get

$$U_j(x_i^{(i)}, y_1^i) = \int \beta_j\big(x_1^{(1)}, \dots, x_n^{(1)}\big)^2 \, \check{\Pi}_i(dx_1^{(1)}, \dots, dx_n^{(i)} | x_i^{(i)}, y_1^{i-1}),$$

where $\check{\Pi}_i(\cdot \,| x_i^{(i)}, y_1^{i-1})$ has already been defined as the law of

$$\big(X^{(1)}, \dots, X^{(i-1)}, (X_{i+1}^{(i)}, \dots, X_n^{(i)})\big)$$

given

$$X_i^{(i)} = x_i^{(i)}, \qquad Y_1^i = y_1^i.$$

Recall that this law is independent of y_i. Therefore, we will write

$$U_j(x_i^{(i)}, y_1^i) = U_j(x_i^{(i)}, y_1^{i-1}).$$

By the property (2.36) of the measure Σ_{i+1}^n, we have

$$T_j(x_i^{(i)}, y_1^i) \le a_j(y_1^{i-1}, x_i^{(i)}, y_i).$$

From inequality (2.40), we get that for every $y_1, \dots, y_n, x_i^{(i)}$ in $\mathbb{R}$,

$$T_j(x_i^{(i)}, y_1^i) \le (\gamma_i^j)^2 \mathbf{1}_{x_i^{(i)} \neq y_i}.$$

From inequality (2.42), we get that for almost every $y_1^{i-1}, y_i, x_i^{(i)}$ with respect to the measure $L_i(dx_i^{(i)}, dy_i | y_1^{i-1}) \, H_1^{i-1}(dy_1, \dots, dy_{i-1})$,

$$T_j(x_i^{(i)}, y_1^i) \le (\tilde{\gamma}_i^j)^2 \mathbf{1}_{x_i^{(i)} \neq y_i}.$$

The end of the proof is identical for γ_i^j and $\tilde{\gamma}_i^j$. We have

$$D_j^{(i)} \le \gamma_i^j \iint \big(U_j(x_i^{(i)}, y_1^{i-1})\big)^{1/2} \mathbf{1}_{x_i^{(i)} \neq y_i} L_i(dx_i^{(i)}, dy_i | y_1^{i-1})$$

$$\times H_1^{i-1}(dy_1, \dots, dy_{i-1}).$$

Integrating with respect to y_i, it follows that

$$D_j^{(i)} \le \gamma_i^j \iint \big(U_j(x_i^{(i)}, y_1^{i-1})\big)^{1/2}\big[g_i(x_i^{(i)}|y_1^{i-1}) - h_i(x_i^{(i)}|y_1^{i-1})\big]_+ d\mu_i(x_i^{(i)}) \times H_1^{i-1}(dy_1, \dots, dy_{i-1}).$$

Then, by the Cauchy–Schwarz inequality,

$$D_j^{(i)} \le \gamma_i^j \int \left(\int U_j(x_i^{(i)}, y_1^{i-1})\, G_i(dx_i^{(i)}|y_1^{i-1})\right)^{1/2} \times \left(\int \left[1 - \frac{h_i(x_i^{(i)}|y_1^{i-1})}{g_i(x_i^{(i)}|y_1^{i-1})}\right]_+^2 G_i(dx_i^{(i)}|y_1^{i-1})\right)^{1/2} H_1^{i-1}(dy_1, \dots, dy_{i-1}).$$

We conclude the argument as in case of C_j, using inequality (2.33) of Lemma 2 for $d_{\mu_i}(h_i(\cdot\,|y_1^{i-1})|g_i(\cdot|y_1^{i-1}))$. We get in this way

$$D_j^{(i)} \le \gamma_i^j(\widetilde{\Delta}_j)^{1/2}(2E_i)^{1/2}. \tag{2.46}$$

We then deduce (2.38) from (2.46) and (2.45). This ends the proof of Theorem 1. □

3. Deviation inequalities for empirical processes. In this section, $X = (X_1, \dots, X_n)$ is a sample of random variables on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, taking values in some measurable space S. We extend the definition of the mixing coefficients γ_i^j and $\tilde{\gamma}_i^j$ as follows. For every $1 \le i < j \le n$ and for $x_i, y_1, \dots, y_i$ in S, let

$$a_j(y_1^{i-1}, x_i, y_i) = \left\| \mathscr{L}(X_j^n|X_1^{i-1} = y_1^{i-1},\ X_i = x_i) - \mathscr{L}(X_j^n|X_1^{i-1} = y_1^{i-1},\ X_i = y_i)\right\|_{\mathrm{TV}}$$

and

$$\left(\gamma_i^j\right)^2 = \sup_{(x_i, y_i)\in S^2}\ \sup_{y_1^{i-1}\in S^{i-1}} a_j(y_1^{i-1}, x_i, y_i).$$

Similarly, for every $1 \le i < j \le n$, let

$$\tilde{a}_j(y_1^i) = \left\|\mathscr{L}(X_j^n|X_1^i = y_1^i) - \mathscr{L}(X_j^n)\right\|_{\mathrm{TV}}$$

and

$$\left(\tilde{\gamma}_i^j\right)^2 = 2 \operatorname*{ess\,sup}_{y_1^i\in S^i,\ \mathscr{L}(X_1^i)} \tilde{a}_j(y_1^i).$$

As previously, we are interested in samples $X = (X_1, \dots, X_n)$ for which $\|\Gamma\|$ may be bounded independently of n, the size of the sample X. Such samples have already been described in Section 2.

Let $\mathscr{F}$ be a countable class of bounded measurable functions g on S. Our aim is to give some exponential deviation inequalities for the supremum of empirical processes. To this task, let

$$Z = \sup_{g\in\mathscr{F}} \left| \sum_{i=1}^{n} g(X_i) \right|.$$

If $\mathscr{F}$ is a finite family of nonnegative functions, a quite simple application of Theorem 1 provides the following deviation inequalities for the random variable Z.

THEOREM 2. *Under the previous notation, assume that* $0 \le g \le C$, $g \in \mathscr{F}$. *Then, for every* $t \ge 0$,

$$\mathbb{P}(Z \ge \mathbb{E}(Z) + t) \le \exp\left(- \frac{t^2}{2C\|\Gamma\|^2(\mathbb{E}(Z) + t)} \right) \tag{3.1}$$

and

$$\mathbb{P}(Z \le \mathbb{E}(Z) - t) \le \exp\left(- \frac{t^2}{2C\|\Gamma\|^2\mathbb{E}(Z)} \right). \tag{3.2}$$

For further purposes, observe that (3.1) is equivalent to saying that, for every $t \ge 0$,

$$\mathbb{P}(Z \ge \mathbb{E}(Z) + t) \le \exp\left(- \frac{1}{4\|\Gamma\|^2} \min\left(\frac{t}{C}, \frac{t^2}{C\mathbb{E}(Z)} \right) \right).$$

Inequalities (3.1) and (3.2) give the exact control of the deviation from the mean. This statement extends in this case the result of Talagrand for Gaussian bounds (see [15]). Actually, this theorem is exactly the extension of Theorem 2.1 [7], in case of dependence. However, this result is limited to the supremum of empirical processes over classes of nonnegative functions. This assumption is restrictive and we want to present now some deviation inequalities for which $\mathscr{F}$ is a class of arbitrary bounded functions. Assume that for every real function g in $\mathscr{F}$, $|g| \le C$. Define the random variable V,

$$V^2 = \sum_{i=1}^{n} \sup_{g\in\mathscr{F}} g(X_i)^2.$$

THEOREM 3. *Under the previous notation, for every* $t \ge 0$,

$$\mathbb{P}(Z \ge \mathbb{E}(Z) + t) \le \exp\left(- \frac{1}{8\|\Gamma\|^2} \min\left(\frac{t}{C}, \frac{t^2}{4\mathbb{E}(V^2)} \right) \right) \tag{3.3}$$

and

$$\mathbb{P}(Z \le \mathbb{E}(Z) - t) \le \exp\left(- \frac{1}{8\|\Gamma\|^2} \min\left(\frac{t}{C}, \frac{t^2}{4\mathbb{E}(V^2)} \right) \right). \tag{3.4}$$

In the independent case ($\|\Gamma\| = 1$), Talagrand actually got a much better result. Namely, for every $t \geq 0$,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq K \exp\left(-\frac{1}{K}\frac{t}{C}\log\left(1 + \frac{Ct}{\mathbb{E}(\Sigma^2)}\right)\right), \tag{3.5}$$

where K is a numerical constant and where

$$\Sigma^2 = \sup_{g\in\mathscr{F}} \sum_{i=1}^{n} g^2(X_i).$$

In particular,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq K \exp\left(-\frac{1}{K}\min\left(\frac{t}{C}, \frac{t^2}{\mathbb{E}(\Sigma^2)}\right)\right). \tag{3.6}$$

Clearly $\Sigma^2 \leq V^2$ so that our results are less precise on this side. It is an open question to prove (3.3) and (3.4), with Σ^2 instead of V^2. Let us recall that in the independent case, for the bound of $\mathbb{P}(Z \geq \mathbb{E}(Z) + t)$ above the mean, [7] and then [12] present an efficient simple proof of (3.5) based on log-Sobolev method. Note also that in the independent case, $\mathbb{E}(\Sigma^2)$ may be bounded by $C\,\mathbb{E}(Z)$ and the supremum of the variances which are then of direct interest in applications (see [16], [7]).

PROOF OF THEOREM 2. By homogeneity, it is enough to deal with the case $C = 1$. Assume $\mathscr{F}$ is a finite class of positive measurable functions,

$$\mathscr{F} = \mathscr{F}_N = \{g_1, \dots, g_N\}.$$

We will prove Theorem 2 for $\mathscr{F} = \mathscr{F}_N$. The result of countable classes of positive measurable functions will follow by monotone convergence. Let

$$f_N(x_1, \dots, x_n) = \max_{1\leq k\leq N} \sum_{i=1}^{n} g_k(x_i).$$

For $1 \leq k \leq N$ and for every $x_1, \dots, x_n$ in S, define

$$\alpha_k(x_1, \dots, x_n) = \begin{cases} 1, & \text{if } k = \inf\left\{1 \leq l \leq N,\ f_N(x_1, \dots, x_n) = \left|\sum_{i=1}^{n} g_l(x_i)\right|\right\}, \\ 0, & \text{otherwise.} \end{cases}$$

According to this definition,

$$f_N(x_1, \dots, x_n) = \sum_{k=1}^{N}\sum_{i=1}^{n} \alpha_k(x_1, \dots, x_n) g_k(x_i)$$

and

$$\sum_{k=1}^{N} \alpha_k(x_1, \dots, x_n) = 1.$$

Let us observe that for all $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ in S^n,

$$f_N(y) - f_N(x) \le \sum_{k=1}^{N} \sum_{i=1}^{n} (g_k(y_i) - g_k(x_i))\alpha_k(y).$$

Let $\alpha(x)$ denote the vector $(\alpha_1(x), \dots, \alpha_N(x))$. For every x in S^n, $\alpha(x)$ is one of the basis elements of $\mathbb{R}^N$. Actually $\alpha(x)$ is the derivative of the supremum norm on $\mathbb{R}^N$. Let $g(x_i) = (g_1(x_i), \dots, g_N(x_i))$, $g(x_i) \in \mathbb{R}^{+N}$. We have, for every x and y,

$$f_N(y) - f_N(x) \le \sum_{i=1}^{n} \langle \alpha(y), g(y_i) - g(x_i) \rangle.$$

As a consequence, if $\tilde{f}_N = -f_N$, we get, for every x and y,

$$\tilde{f}_N(y) - \tilde{f}_N(x) \le \sum_{i=1}^{n} \langle \alpha(x), g(x_i) - g(y_i) \rangle.$$

Since g_k are nonnegative, it follows that

$$f_N(y) - f_N(x) \le \sum_{i=1}^{n} \langle \alpha(y), g(y_i) \rangle \mathbf{1}_{x_i \ne y_i} \tag{3.7}$$

and

$$\tilde{f}_N(y) - \tilde{f}_N(x) \le \sum_{i=1}^{n} \langle \alpha(x), g(x_i) \rangle \mathbf{1}_{x_i \ne y_i}. \tag{3.8}$$

From this stage, the proof is similar to the proof of Corollary 3. P is the law of $(X_1, \dots, X_n)$ on S^n. Let Q be a probability measure on S^n with density g with respect to P. For every measure Π on $S^n \times S^n$ with marginals Q and P, that is, $\Pi \in \mathscr{M}(P, Q)$,

$$\int f_N(y)\,dQ(y) - \int f_N(x)\,dP(x) = \iint (f_N(y) - f_N(x))\,d\Pi(x, y).$$

Therefore, by (3.7),

$$\int f_N(y)\,dQ(y) - \int f_N(x)\,dP(x) \le \iint \sum_{n=1}^{n} \langle \alpha(y), g(y_i) \rangle \mathbf{1}_{x_i \ne y_i}\,d\Pi(x, y).$$

Integrating with respect to the variable x and then using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
&\int f_N(y)\,dQ(y) - \int f_N(x)\,dP(x) \\
&\quad \le \left[\int \sum_{i=1}^{n} \langle \alpha(y), g(y_i) \rangle^2\,dQ(y) \right]^{1/2} \left[\int \sum_{n=1}^{n} \mathbb{P}^2(X_i \ne y_i | Y_i = y_i)\,dQ(y) \right]^{1/2},
\end{aligned}$$

where (X, Y) denotes a pair of random variable taking values in $\mathbb{R}^n \times \mathbb{R}^n$, and with law Π. According to the definition of $d_2(P, Q)$, minimizing the right-hand side over all measures Π in $\mathscr{M}(P, Q)$ yields

$$\int f_N(y)\,dQ(y) - \int f_N(x)\,dP(x) \le \left[\int \sum_{i=1}^n \langle \alpha(y), g(y_i)\rangle^2\,dQ(y)\right]^{1/2} d_2(P, Q).$$

Similarly, for the function $\tilde{f}_N$, we get from (3.8),

$$\int \tilde{f}_N(y)\,dQ(y) - \int \tilde{f}_N(x)\,dP(x) \le \left[\int \sum_{i=1}^n \langle \alpha(x), g(x_i)\rangle^2\,dP(x)\right]^{1/2} d_2(Q, P).$$

Let us now define h_N^2 on S^n by

$$h_N^2(x) = \sum_{i=1}^n \langle \alpha(x), g(x_i)\rangle^2, \qquad x \in S^n.$$

With our previous notation, $h_N^2(X) = \Sigma^2$. Applying (2.11) of Theorem 1, we get

$$\int f_N\,dQ - \int f_N\,dP \le \sqrt{2\|\Gamma\|^2 E_Q(h_N^2)\operatorname{Ent}_P\left(\frac{dQ}{dP}\right)}.$$

Therefore, as in the proof of Corollary 3, for every $\lambda > 0$,

$$\int f_N g\,dP - \int f_N\,dP \le \lambda\frac{\|\Gamma\|^2 E_Q(h_N^2)}{2} + \frac{1}{\lambda}\operatorname{Ent}_P(g).$$

Finally, for every $\lambda > 0$,

$$\int \left[\lambda(f_N - E_P(f_N)) - \lambda^2\frac{\|\Gamma\|^2 h_N^2}{2}\right] g\,dP \le \operatorname{Ent}_P(g). \tag{3.9}$$

For the function $\tilde{f}_N$, the result is quite different. For every $\lambda > 0$,

$$\int \left[\lambda(\tilde{f}_N - E_P(\tilde{f}_N)) - \lambda^2\frac{\|\Gamma\|^2 E_P(h_N^2)}{2}\right] g\,dP \le \operatorname{Ent}_P(g). \tag{3.10}$$

The exponential inequalities then follow with a good choice for the density g. From the inequality (3.9) we get

$$\int \exp\left[\lambda(f_N - E_P(f_N)) - \lambda^2\frac{\|\Gamma\|^2 h_N^2}{2}\right] dP \le 1. \tag{3.11}$$

Similarly from (3.10), we get

$$\int \exp\left[\lambda\big(\tilde{f}_N - E_P(\tilde{f}_N)\big)\right] dP \le \exp\left(\lambda^2\frac{\|\Gamma\|^2 E_P(h_N^2)}{2}\right). \tag{3.12}$$

Recall that

$$Z = f_N(X) = -\tilde{f}_N(X)$$

and that

$$\Sigma^2 = h_N^2(X).$$

If, for every $1 \leq k \leq N$, $0 \leq g_k \leq 1$, then, $\Sigma^2 \leq Z$. Thus from the exponential inequality (3.11) we get that for every $\lambda > 0$,

$$\mathbb{E}\left(\exp\left[Z\lambda\left(1 - \frac{\|\Gamma\|^2\lambda}{2}\right) - \lambda\mathbb{E}(Z)\right]\right) \leq 1, \tag{3.13}$$

and from (3.12) we get, for every $\lambda > 0$,

$$\mathbb{E}\big(\exp[-\lambda(Z - \mathbb{E}(Z))]\big)\,dP \leq \exp\left(\lambda^2 \frac{\|\Gamma\|^2\mathbb{E}(Z)}{2}\right). \tag{3.14}$$

Using inequality (3.13) we get, by Chebyshev's inequality, that for every $0 \leq \lambda \leq 2/\|\Gamma\|^2$, and for every $t \geq 0$,

$$\mathbb{P}(Z \geq \mathbb{E}(Z) + t) \leq \exp\left[-t\lambda\left(1 - \frac{\|\Gamma\|^2\lambda}{2}\right) + \lambda^2\frac{\|\Gamma\|^2\mathbb{E}(Z)}{2}\right].$$

Choose then

$$\lambda = \frac{t}{\|\Gamma\|^2(t + \mathbb{E}(Z))},$$

and (3.1) follows. From (3.14), we get in the same way that for every $\lambda \geq 0$ and every $t \geq 0$,

$$\mathbb{P}(-Z \geq -\mathbb{E}(Z) + t) \leq \exp\left[-t\lambda + \lambda^2\frac{\|\Gamma\|^2\mathbb{E}(Z)}{2}\right].$$

Optimizing in λ yields the deviation inequality (3.2). The proof of Theorem 2 is thus complete. □

We now present the proof of Theorem 3.

PROOF OF THEOREM 3. As for the proof of Theorem 2, we may assume that $\mathscr{F}$ is finite. Let

$$f_N(x_1, \ldots, x_n) = \max_{1\leq k\leq n}\left|\sum_{i=1}^{n} g_k(x_i)\right|.$$

For $1 \leq k \leq N$ and for every $x_1, \ldots, x_n$ in S, define

$$\alpha_k(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } k = \inf\left\{1 \leq l \leq N,\ f_N(x_1, \ldots, x_n) = \left|\sum_{i=1}^{n} g_l(x_i)\right|\right\}, \\ 0, & \text{otherwise.} \end{cases}$$

According to this definition,

$$f_N(x_1,\dots,x_n)=\sum_{k=1}^N \alpha_k(x_1,\dots,x_n)\left|\sum_{i=1}^n g_k(x_i)\right|$$

and

$$\sum_{K=1}^N \alpha_k(x_1,\dots,x_n)=1.$$

Let us observe that for all $x=(x_1,\dots,x_n)$ and $y=(y_1,\dots,y_n)$ in S^n,

$$f_N(y)-f_N(x)\le \sum_{k=1}^N \alpha_k(y)\left|\sum_{i=1}^n (g_k(y_i)-g_k(x_i))\right|.$$

By the triangle inequality,

$$\left|g_k(y_i)-g_k(x_i)\right|\le \left|g_k(y_i)\right|\mathbf{1}_{x_i\ne y_i}+\left|g_k(x_i)\right|\mathbf{1}_{x_i\ne y_i}.$$

Therefore, if $|g(x_i)|$ denotes the vector $(|g_1(x_i)|,\dots,|g_N(x_i)|)$, we have

$$f_N(y)-f_N(x)\le \sum_{i=1}^n \big\langle \alpha(y),|g(y_i)|\big\rangle \mathbf{1}_{x_i\ne y_i}+\sum_{i=1}^n \big\langle \alpha(y),|g(x_i)|\big\rangle \mathbf{1}_{x_i\ne y_i}.$$

Bounding $\langle \alpha(y),|g(x_i)|\rangle$, by $\max_{1\le k\le N}|g_k(x_i)|$, it follows that

$$f_N(y)-f_N(x)\le \sum_{i=1}^n \big\langle \alpha(y),|g(y_i)|\big\rangle \mathbf{1}_{x_i\ne y_i}+\sum_{i=1}^n \max_{1\le k\le N}|g_k(x_i)|\mathbf{1}_{x_i\ne y_i}. \tag{3.15}$$

Let Q be a probability measure on S^n with density g with respect to P. For every measure Π in $\mathscr{M}(P,Q)$,

$$\int f_N(y)\,dQ(y)-\int f_N(x)\,dP(x)=\int\!\!\int (f_N(y)-f_N(x))\,d\Pi(x,y).$$

Then, by the Cauchy–Schwarz inequality, from (3.15) we get

$$\begin{aligned}
&\int f_N(y)\,dQ(y)-\int f_N(x)\,dP(x)\\
&\quad\le\left[\int \sum_{i=1}^n \langle \alpha(y),g(y_i)\rangle^2\,dQ(y)\right]^{1/2}\left(\int \sum_{i=1}^n \mathbb{P}^2(X_i\ne y_i|Y_i=y_i)\,dQ(y)\right)^{1/2}\\
&\qquad+\left[\int \sum_{i=1}^n \max_{1\le k\le N} g_k^2(x_i)\,dP(x)\right]^{1/2}\left(\int \sum_{i=1}^n \mathbb{P}^2(Y_i\ne x_i|X_i=x_i)\,dP(x)\right)^{1/2},
\end{aligned}$$

where (X,Y) denotes a pair of random variables taking values in $\mathbb{R}^n\times\mathbb{R}^n$ whose law is Π. From the proof of Theorem 1, we know that there exists a measure Π in $\mathscr{M}(P,Q)$ with

$$\left(\int \sum_{i=1}^n \mathbb{P}^2(X_i\ne y_i|Y_i=y_i)\,dQ(y)\right)^{1/2}\le \|\Gamma\|\sqrt{2\,\mathrm{Ent}_P\left(\frac{dQ}{dP}\right)} \tag{3.16}$$

and

$$\left(\int \sum_{i=1}^{n} \mathbb{P}^2(Y_i \neq x_i | X_i = x_i)\, dP(x)\right)^{1/2} \leq \|\Gamma\| \sqrt{2 \operatorname{Ent}_P\left(\frac{dQ}{dP}\right)}. \tag{3.17}$$

Recall the definition of h_N^2,

$$h_N^2(x) = \sum_{i=1}^{n} \langle \alpha(x), g(x_i)\rangle^2, \qquad x \in S^n.$$

For every x in S^n, let

$$l_N^2(x) = \sum_{i=1}^{n} \max_{1 \leq k \leq N} g_k^2(x_i).$$

Choosing the measure Π satisfying (3.16) and (3.17), we get that

$$\begin{aligned} &\int f_N(y)\, dQ(y) - \int f_N(x)\, dP(x) \\ &\quad \leq \sqrt{2\|\Gamma\|^2 E_Q(h_N^2) \operatorname{Ent}_P\left(\frac{dQ}{dP}\right)} + \sqrt{2\|\Gamma\|^2 E_P(l_N^2) \operatorname{Ent}_P\left(\frac{dQ}{dP}\right)}. \end{aligned}$$

Therefore, using the same argument as in the proof of Theorem 2, for every $\lambda > 0$,

$$\begin{aligned} &\int f_N(y)\, dQ(y) - \int f_N(x)\, dP(x) \\ &\quad \leq \lambda \frac{\|\Gamma\|^2}{2} (\mathrm{E}_Q(h_N^2) + \mathrm{E}_P(l_N^2)) + \frac{2}{\lambda} \operatorname{Ent}_P(g). \end{aligned}$$

Finally, for every $\lambda > 0$,

$$\int \left[\frac{\lambda}{2}(f_N - \mathrm{E}_P(f_N)) - \lambda^2 \frac{\|\Gamma\|^2}{4} (h_N^2 + \mathrm{E}_P(l_N^2))\right] g\, dP \leq \operatorname{Ent}_P(g).$$

With a good choice for the density g, we get that

$$\int \exp\left[\frac{\lambda}{2}(f_N - \mathrm{E}_P(F_N)) - \lambda^2 \frac{\|\Gamma\|^2}{4} (h_N^2 + \mathrm{E}_P(l_N^2))\right] dP \leq 1.$$

Since $Z = f_N(X)$, $\Sigma^2 = h_N^2(X)$ and $V^2 = l_N^2(X)$, we have

$$\int \exp\left[\frac{\lambda}{2}(Z - \mathbb{E}(Z)) - \lambda^2 \frac{\|\Gamma\|^2}{4} (\Sigma^2 + \mathbb{E}(V^2))\right] d\mathbb{P} \leq 1.$$

By the Cauchy–Schwarz inequality, it follows that

$$\int \exp\left(\frac{\lambda}{4}(Z - \mathbb{E}(Z))\right) dP \leq \left[\int \exp\left(\lambda^2 \frac{\|\Gamma\|^2}{4} \Sigma^2\right) dP\right]^{1/2} \exp\left(\lambda^2 \frac{\|\Gamma\|^2}{8} \mathbb{E}(V^2)\right).$$

Applying inequality (3.13) to the random variable Σ^2 yields that, for every $0 \le \mu \le 1/C^2\|\Gamma\|^2$,

$$\mathbb{E}\biggl(\exp\biggl(\frac{\mu}{2}\Sigma^2\biggr)\biggr) \le \exp\bigl(\mu\mathbb{E}(\Sigma^2)\bigr).$$

Choosing $\mu = \lambda^2(\|\Gamma\|^2/2)$, we get for every $0 \le \lambda \le 1/C\|\Gamma\|^2$,

$$\int \exp\biggl(\lambda^2\frac{\|\Gamma\|^2}{4}\Sigma^2\biggr)\,dP \le \exp\biggl(\lambda^2\frac{\|\Gamma\|^2}{2}\mathbb{E}(\Sigma^2)\biggr).$$

Finally, for every $0 \le \lambda \le 1/C\|\Gamma\|^2$,

$$\int \exp\biggl(\frac{\lambda}{4}(Z-\mathbb{E}(Z))\biggr)\,dP \le \exp\biggl(\lambda^2\frac{\|\Gamma\|^2}{4}(\mathbb{E}(V^2)+\mathbb{E}(\Sigma^2))\biggr) \le \exp\biggl(\lambda^2\frac{\|\Gamma\|^2}{2}\mathbb{E}(V^2)\biggr).$$

Then the proof of (3.3) is easily completed by Chebyshev's inequality. The proof of (3.4) is identical to the one of (3.3). This ends the proof of Theorem 3. □

Acknowledgment. I thank my academic adviser Michel Ledoux for presenting and explaining to me in recent years his work on concentration inequalities that finally led to this study. I am grateful to him for his enlightening comments and for his careful reading of the manuscript.

REFERENCES

[1] BOBKOV, S. (1996). Some extremal properties of Bernouilli distribution. *Theory Probab. Appl.* **41** 877–884.

[2] BOBKOV, S. and GÖTZE, F. (1997). Exponential integrability and transportation cost under logarithmic Sobolev inequalities. *J. Funct. Anal.* To appear.

[3] DEMBO, A. (1997). Information inequalities and concentration of measure. *Ann. Probab.* **25** 927–939.

[4] DEMBO, A. and ZEITOUNI, O. (1996). Transportation approach to some concentration inequalities in product spaces. *Elect. Comm. Probab.* **1**.

[5] DOUKHAN, P. (1995). *Mixing: Properties and Examples. Lecture Notes in Statist.* **85**. Springer, Berlin.

[6] FIEBIG, D. (1993). Mixing Properties of a class of Bernoulli processes. *Trans. Amer. Math. Soc.* **338** 479–492.

[7] LEDOUX, M. (1996). On Talagrand's deviation inequalities for product measure. *ESAIM Probab. Statist.* **1** 63–87.

[8] MARTON, K. (1996). A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6** 556–571.

[9] MARTON, K. (1997). A measure concentration inequality for contracting Markov chains Erratum. *Geom. Funct. Anal.* **7** 609–613.

[10] MARTON, K. (1998). Measure concentration for a class of random processes. *Probab. Theory Related Fields* **110** 427–429.

[11] MARTON, K. (1999). On a measure concentration inequality of Talagrand for dependent random variables. Preprint.

[12] MASSART, P. (1998). About the constants in Talagrand's deviation inequalies for empirical processes. Preprint.

[13] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, New York.

[14] MILMAN, V. D. and SCHECHTMAN, G. (1986). Asymptotic theory of finite dimensional normed spaces. *Lecture Notes in Math.* **1200**. Springer, Berlin.

[15] TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. I.H.E.S.* **81** 73–205.
[16] TALAGRAND, M. (1996). A new look at independence. *Ann. Probab.* **24** 1–34.
[17] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563.

LABORATOIRE DE STATISTIQUES ET PROBABILITÉS
UNIVERSITÉ PAUL SABATIER
118 ROUTE DE NARBONNE
TOULOUSE
FRANCE
E-mail: samson@cict.fr