The Annals of Probability
1997, Vol. 25, No. 2, 927–939

INFORMATION INEQUALITIES AND CONCENTRATION OF MEASURE[1]

BY AMIR DEMBO

Technion–Israel Institute of Technology

We derive inequalities of the form $\Delta(P, Q) \le H(P|R)+H(Q|R)$ which hold for every choice of probability measures P, Q, R, where $H(P|R)$ denotes the relative entropy of P with respect to R and $\Delta(P, Q)$ stands for a coupling type "distance" between P and Q. Using the chain rule for relative entropies and then specializing to Q with a given support we recover some of Talagrand's concentration of measure inequalities for product spaces.

1. Introduction. In [9], Talagrand provided a variety of concentration of measure inequalities which apply in every product space Ω^N equipped with a product (probability) measure R. These inequalities are extremely useful in combinatorial applications such as the longest common/increasing subsequence, in statistical physics applications such as the study of spin glass models and in areas touching upon functional analysis, such as probability in Banach spaces (cf. [9]–[11] and the references therein). For suitably chosen "distance" functions $f(\cdot)$, these inequalities are of the form

$$\int \exp(tf(A_1, \dots, A_q, \mathrm{x}))\, d\mathrm{R}(\mathrm{x}) \le \exp(C(t, \alpha)) \prod_{i=1}^{q} \mathrm{R}(A_i)^{-\alpha}, \tag{1}$$

for some constants $q \in \mathbb{N}$, $\alpha, t > 0$ and $C(t, \alpha) < \infty$, and hold for every (measurable) $A_i \subset \Omega^N$, $i = 1, \dots, q$. Of most interest are the "dimension-free" inequalities in which q, α, t are independent of N and $C(t, \alpha) = 0$. Not to be distracted from the main course of this paper, we follow Talagrand's convention and hereafter ignore all measurability questions (these can either be taken care of by considering upper integrals and outer probabilities or circumvented by assuming Ω is Polish, the A_i are compact and all probability measures encountered are Borel measures). Three "distance" functions that play a prominent role in [9] are the "control by q points"

$$f_q(A_1, \dots A_q, \mathrm{x}) = \inf_{\substack{\mathrm{y}^i \in A_i \\ i=1,\dots,q}} \sum_{k=1}^{N} 1_{x_k \notin \{y^i_k,\, i=1,\dots,q\}},$$

the "penalties"

$$f_h(A, \mathrm{x}) = \inf_{\mathrm{y} \in A} \sum_{k=1}^{N} h(x_k, y_k)$$

Received November 1995; revised September 1996.

[1]Partially supported by NSF Grant DMS-94-03553 and by a US–Israel BSF grant.

AMS 1991 *subject classifications*. 60E15, 28A35.

Key words and phrases. Concentration of measure, information inequalities.

for $h\colon \Omega\times\Omega\to[0,\infty)$ and the "convex hull"

$$f_\alpha(A,\mathsf{x})=\inf_{\mathsf{s}\in V_A(\mathsf{x})}\sum_{k=1}^N \xi_\alpha(s_k),$$

where $V_A(\mathsf{x})\stackrel{\triangle}{=}\text{conv-hull}\{(1_{x_1\neq y_1},\dots,1_{x_N\neq y_N})\colon \mathsf{y}\in A\}$, $\alpha>0$ and $\xi_\alpha\colon[0,1]\to[0,\infty)$ is such that

$$\xi_\alpha(u)\le\alpha(1-u)\log(1-u)-(\alpha+1-\alpha u)\log\left(\frac{\alpha+1-\alpha u}{1+\alpha}\right)\qquad\forall u\in[0,1]. \tag{2}$$

The proofs of the inequalities of the form (1) provided in [9] are all based on an induction on N, the key step of which is to fix $x_{N+1}=\omega$ and then apply the induction hypothesis for the N dimensional sets $A(\omega)=\{(y_1,\dots,y_N)\colon (y_1,\dots,y_N,\omega)\in A\}$ and $B=\{(y_1,\dots,y_N)\colon (y_1,\dots,y_N,z)\in A$ for some $z\}$.

Marton, in [6], building upon [5], proposed a new approach to concentration inequalities, based on the use of information inequalities, and in [6, 7] applied this approach to extend some of Talagrand's results to the context of contracting Markov chains.

Marton's work is the impetus for this paper, in which we concentrate on the case of product measures and recover the sharper variants of the inequalities of [9] (see the discussion following Theorem 1 below).

Specifically, with $\mathscr{M}_N(\mathsf{Q}_1,\dots,\mathsf{Q}_q,\mathsf{P})$ denoting the set of all probability measures on $(\Omega^N)^{q+1}$ whose marginals are the prescribed probability measures $\mathsf{Q}_1,\dots,\mathsf{Q}_q,\mathsf{P}$ on Ω^N, we consider the following coupling type "distances" between probability measures in a product space Ω^N:

$$d_q(\mathsf{Q}_1,\dots,\mathsf{Q}_q,\mathsf{P})=\inf_{\pi\in\mathscr{M}_N(\mathsf{Q}_1,\dots,\mathsf{Q}_q,\mathsf{P})}\sum_{k=1}^N\pi(X_k\notin\{Y_k^i,\ i=1,\dots,q\}),$$

$$d_h(\mathsf{Q},\mathsf{P})=\inf_{\pi\in\mathscr{M}_N(\mathsf{Q},\mathsf{P})}\sum_{k=1}^N\int h(x_k,y_k)\,d\pi(\mathsf{x},\mathsf{y}),$$

$$d_\alpha(\mathsf{Q},\mathsf{P})=\inf_{\pi\in\mathscr{M}_N(\mathsf{Q},\mathsf{P})}\sum_{k=1}^N\int \xi_\alpha(\pi(Y_k\neq X_k|\mathsf{X}=\mathsf{x}))\,d\mathsf{P}(\mathsf{x}),$$

where $(\mathsf{Y}^1,\dots,\mathsf{Y}^q,\mathsf{X})$ has the joint law π and X_k (Y_k^i) denotes the Ω-valued kth marginal of X (Y^i, respectively).

Recall that the relative entropy of P with respect to R is

$$H(\mathsf{P}|\mathsf{R})=\begin{cases}\displaystyle\int\frac{d\mathsf{P}}{d\mathsf{R}}\log\frac{d\mathsf{P}}{d\mathsf{R}}\,d\mathsf{R}, & \displaystyle\frac{d\mathsf{P}}{d\mathsf{R}}\text{ exists},\\ \infty, & \text{otherwise}.\end{cases}$$

Theorem 1, our main result (whose proof is provided in Section 2), states that for appropriate choices of $\alpha,t>0$ and $C(t,\alpha)<\infty$ the functionals d_q, d_h and

d_α satisfy the inequality

$$td_\bullet(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}) \le H(\mathsf{P}|\mathsf{R}) + \alpha \sum_{i=1}^{q} H(\mathsf{Q}_i|\mathsf{R}) + C(t, \alpha) \tag{3}$$

for every choice of probability measures P, Q_i and every product measure R.

THEOREM 1. *Suppose* $\mathsf{R} = \prod_k R_k$ *is a product measure on* Ω^N.

(i) *The functional* d_α *satisfies inequality* (3) *with* $C(t,\alpha) = 0$ *for* $t = 1$, *any* $\alpha > 0$ *and* ξ_α *satisfying* (2).

(ii) *Inequality* (3) *holds for* $d_q, q > 1$, $C(t,\alpha) = 0$ *and* t *which is the unique positive solution of* $1 + \alpha q = e^t + \alpha q e^{-t/\alpha}$. *Moreover*, $d_q, q = 1$, *satisfies* (3) *for any* $t, \alpha > 0$, *but now with* $C(t, \alpha) = N \log a(t, \alpha)$, *where*

$$a(t,\alpha) = \frac{\alpha^\alpha}{(1+\alpha)^{1+\alpha}} \frac{(e^t - e^{-t/\alpha})^{1+\alpha}}{(1 - e^{-t/\alpha})(e^t - 1)^\alpha}. \tag{4}$$

(iii) *The functional* d_h *satisfies inequality* (3) *for* Ω *a Polish space and* $h \in B(\Omega \times \Omega)$, $\alpha = 1$, *any* $t > 0$ *and* $C(t, \alpha) = \sum_{k=1}^{N} b(t, h, R_k)$ *with*

$$b(t, h, R) = \sup_{\substack{g \ge 0 \\ \hat{g}(x) \le t h(x,y) + g(y)}} \log\left(\int e^{\hat{g}}\, dR \int e^{-g}\, dR\right). \tag{5}$$

We next show that some of the concentration of measure inequalities of [9] are direct corollaries of Theorem 1. To this end observe that the inequality

$$\int f_\bullet(A_1, \dots, A_q, \mathsf{x})\, d\mathsf{P}(\mathsf{x}) \le d_\bullet(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}) \tag{6}$$

holds for every P and Q_i such that $\operatorname{supp} \mathsf{Q}_i \subseteq A_i$ when considering d_q, d_h and d_α paired with f_q, f_h and f_α, respectively.

The following simple lemma shows that whenever (6) holds, the inequality (1) is a consequence of (3) [for the same values of α, t and $C(t, \alpha)$].

LEMMA 1. *Suppose that for a probability measure* R *and some* $q \in \mathbb{N}$, $\alpha, t > 0$ *and* $C(t,\alpha) < \infty$, *the inequality (3) holds for every choice of probability measures* P, Q_i. *Then (1) holds provided that inequality (6) holds for every choice of* P *and* Q_i *such that* $\operatorname{supp} \mathsf{Q}_i \subseteq A_i$.

PROOF. Set $\mathsf{Q}_i = \mathsf{R}(\cdot|A_i)$ for which $H(\mathsf{Q}_i|\mathsf{R}) = \log(1/(\mathsf{R}(A_i)))$ and P_m is such that

$$\frac{d\mathsf{P}_m}{d\mathsf{R}} = \frac{\exp(tf(A_1, \dots, A_q, \mathsf{x}) \wedge m)}{\int \exp(tf(A_1, \dots, A_q, \mathsf{x}) \wedge m)\, d\mathsf{R}(\mathsf{x})}.$$

Then, by (6) and evaluation of $H(\mathsf{P}_m|\mathsf{R})$,

$$\log \int \exp(tf(A_1, \dots A_q, \mathsf{x}) \wedge m)\, d\mathsf{R}(\mathsf{x}) \le td(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}_m) - H(\mathsf{P}_m|\mathsf{R}).$$

Hence, (3) implies that

$$\int \exp(tf(A_1, \dots A_q, \mathsf{x}) \wedge m)\, d\mathsf{R}(\mathsf{x}) \le \exp(C(t,\alpha)) \prod_{i=1}^{q} \mathsf{R}(A_i)^{-\alpha}.$$

Letting $m \to \infty$ we obtain (1) by monotone convergence. □

REMARK 1. Conversely to Lemma 1, if for some $q \in \mathbb{N}$, $f(\cdot)$ bounded, Ω Polish, $\alpha, t > 0$ and a collection $\mathscr{R}$ of probability measures on Ω^N,

$$\exp(C(t,\alpha)) = \sup_{\substack{\mathsf{R}\in\mathscr{R} \\ A_i \subseteq \Omega^N}} \prod_{i=1}^{q} \mathsf{R}(A_i)^{\alpha} \int \exp(tf(A_1, \dots, A_q, \mathsf{x}))\, d\mathsf{R}(\mathsf{x}) \tag{7}$$

for some $C(t,\alpha) < \infty$, then (6) holds for

$$d(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}) = t^{-1}\Big[C(t,\alpha) + \inf_{\mathsf{R}\in\mathscr{R}} \alpha \sum_{i=1}^{q} H(\mathsf{Q}_i|\mathsf{R}) + H(\mathsf{P}|\mathsf{R})\Big] \tag{8}$$

and every choice of P and Q_i such that $\operatorname{supp} \mathsf{Q}_i \subseteq A_i$.

Indeed, fixing $A_i \subseteq \Omega^N$ we have by (8) that

$$\begin{aligned}
&\inf_{\mathsf{P}, \{\mathsf{Q}_i:\ \operatorname{supp}\mathsf{Q}_i \subseteq A_i\}} \Big\{ td(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}) - \int tf(A_1, \dots, A_q, \mathsf{x})\, d\mathsf{P}(\mathsf{x}) \Big\} \\
&\quad = C(t,\alpha) + \inf_{\mathsf{R}\in\mathscr{R}} \Big\{ \alpha \sum_{i=1}^{q} \inf_{\{\mathsf{Q}_i:\ \operatorname{supp}\mathsf{Q}_i \subseteq A_i\}} H(\mathsf{Q}_i|\mathsf{R}) \\
&\qquad\qquad - \sup_{\mathsf{P}} \Big\{ \int tf(A_1, \dots, A_q, \mathsf{x})\, d\mathsf{P}(\mathsf{x}) - H(\mathsf{P}|\mathsf{R}) \Big\} \Big\} \\
&\quad = C(t,\alpha) + \inf_{\mathsf{R}\in\mathscr{R}} \Big\{ -\alpha \sum_{i=1}^{q} \log \mathsf{R}(A_i) - \log \int \exp(tf(A_1, \dots, A_q, \mathsf{x}))\, d\mathsf{R}(\mathsf{x}) \Big\} \\
&\quad \ge 0.
\end{aligned}$$

The inequality in the preceding line is due to (7), whereas the equality follows by the Donsker–Varadhan formula (cf. [1], (6.2.14)) and the well known inequality $H(\mathsf{Q}|\mathsf{R}) \ge -\log \mathsf{R}(\operatorname{supp} \mathsf{Q})$.

In particular, if (1) holds, it should always be possible to derive it by proving (3) and (6) for an appropriate choice of $d(\cdot)$. Moreover, equality in (3) and (6) for R and the same Q_i and P implies equality in (1) for $A_i = \operatorname{supp} \mathsf{Q}_i$ and R.

REMARK 2. Combining Lemma 1 with part (i) of Theorem 1 yields [9], Theorem 4.2.4, whereas part (ii) yields likewise [9], Theorem 3.1.1, and [9], Proposition 2.2.1 (for $q > 1$ and $q = 1$, respectively). Combining part (iii) of Theorem 1 above and the bounds on $b(t, h, R)$ provided in [9], Propositions 2.4.2 and 2.5.2, we get [9], Theorems 2.4.1 and 2.5.1, for $h \ge 0$ and bounded. The general case then follows by standard approximation arguments. Altogether, Theorem 1 recovers the results of Sections 2.1, 2.2, 2.4, 3.1, 3.2, 4.1 and 4.2 of [9]. The optimality of the rhs of (2) in the context of f_α is observed in [9], Lemma 4.2.1, albeit from a seemingly different reason, related to the induction technique of [9].

Marton [7] combined the coupling characterization of the total variation distance with information inequalities of the form of Pinsker's inequality $\Delta_1(Q, P) \leq \sqrt{H(P|Q)/2}$ to prove that

$$\sqrt{2d_{u^2/4}(\mathrm{Q}, \mathrm{P})} \leq \sqrt{H(\mathrm{P}|\mathrm{R})} + \sqrt{H(\mathrm{Q}|\mathrm{R})}, \tag{9}$$

leading to [9], Corollary 4.2.5 [with $d_{u^2/4}(\cdot,\cdot)$ denoting d_α for $\alpha = 1$ and $\xi_\alpha(u) = u^2/4$]. Note that (9) trivially holds when $d_{u^2/4}(\mathrm{Q}, \mathrm{P}) \leq H(\mathrm{Q}|\mathrm{R})/2$. Thus, considering part (ii) of Theorem 1 for $\xi_\alpha(u) = \alpha u^2/(2(\alpha+1))$ (which satisfies (2), cf. [9], Lemma 4.2.2) and $\alpha = \sqrt{2d_{u^2/4}(\mathrm{Q}, \mathrm{P})/H(\mathrm{Q}|\mathrm{R})} - 1 \geq 0$, we also recover (9).

Our proof of Theorem 1 uses the extended coupling of Proposition 2(i) to handle the case of d_q and the well known Lemma 4 to handle that of d_h. To establish a result of the form (3) for a variety of "distance" functionals, with sharper constants, our proof also differs from Marton's in deriving in Proposition 1 the "linearized" information inequalities of the form (11) which might be of some independent interest.

We note in passing that [10] contains new concentration inequalities for product spaces which are possibly sharper than those in [9]. The proofs in [10] are again by means of the basic induction alluded to above. In Section 3 we outline how [10], Theorem 5.4, follows from an extension of part (ii) of Theorem 1, whereas in [2] we use a different variant of the "transportation method" to recover [10], Theorem 2.1, apart from the exact value of certain constants. It is yet unclear to what extent one may recover or even improve upon the inequalities of [10], Theorems 3.1, 4.2 and 5.1, by using the transportation method.

Talagrand [11] used the "transportation method" with a different coupling than the one used here (compare Lemma 4 with [11], (2.1)), showing that (3) holds for d_h with $C(t, \alpha) = 0$, $t = \alpha = q = 1$ and Q = R when considering either $h(x, y) = (x-y)^2/2$ and R the standard Gaussian measure or $h(x, y) = (1-b)(b|x-y| - 1 - e^{-b|x-y|})/b$ for some $b < 1$ and R the product of one-dimensional standard Laplace measures.

Ledoux [4] presented a direct derivation of some consequences of the abstract inequalities of [9] and [10] out of Poincaré and logarithmic Sobolev inequalities.

2. Proof of Theorem 1.

The "distance" functionals between probability measures on Ω^N which we consider are of the form

$$d(\mathrm{Q}_1, \ldots, \mathrm{Q}_q, \mathrm{P}) = \inf_{\pi \in \mathscr{M}_N(\mathrm{Q}_1, \ldots, \mathrm{Q}_q, \mathrm{P})} \sum_{k=1}^{N} g_k(\pi). \tag{10}$$

With each such functional we associate a basic "distance" functional $\Delta(Q_1, \dots, Q_q, P)$ such that for some $\alpha, t > 0$, $c(t, \alpha) < \infty$ and every choice of probability measures P, Q_i, R on Ω,

$$t\Delta(Q_1, \dots, Q_q, P) \le H(P|R) + \alpha \sum_{i=1}^{q} H(Q_i|R) + c(t, \alpha)\,. \tag{11}$$

The next lemma obtains the inequality (3) as a consequence of the basic information inequality (11), and is the only place in our proof where we rely on R being a product measure.

LEMMA 2. *Suppose that for every* $\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}$ *and* $\varepsilon > 0$ *there are* $(\mathsf{Y}^1, \dots, \mathsf{Y}^q, \mathsf{X})$ *of some joint law* $\pi = \pi_\varepsilon \in \mathscr{M}_N(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P})$ *such that for* $k = 1, 2, \dots, N$,

$$\begin{aligned} g_k(\pi) \le E\Delta\big(\mathsf{Q}_1(Y_k^1|Y_1^1, \dots, Y_{k-1}^1), \dots, \mathsf{Q}_q(Y_k^q|Y_1^q, \dots, Y_{k-1}^q),& \\ \mathsf{P}(X_k|X_1, \dots, X_{k-1})\big) + \varepsilon& \end{aligned} \tag{12}$$

[where $\mathsf{Q}_i(Y_k^i|Y_1^i, \dots, Y_{k-1}^i)$ *and* $\mathsf{P}(X_k|X_1, \dots, X_{k-1})$ *denote the corresponding regular conditional probability distributions]. Then (11) implies that (3) holds for every product measure* R *on* Ω^N *with* $C(t, \alpha) = Nc(t, \alpha)$.

PROOF. By (10), (11) and (12) we have that for every $\varepsilon > 0$, probability measures Q_i, P on Ω^N and R_k on Ω,

$$\begin{aligned} td(\mathsf{Q}_1, \dots, \mathsf{Q}_q, \mathsf{P}) &\le \sum_{k=1}^{N} Et\Delta\big(\mathsf{Q}_1(Y_k^1|Y_1^1, \dots, Y_{k-1}^1), \dots, \\ &\qquad\qquad \mathsf{P}(X_k|X_1, \dots, X_{k-1})\big) + Nt\varepsilon \\ &\le E\bigg[\sum_{k=1}^{N} H(\mathsf{P}(X_k|X_1, \dots, X_{k-1})|R_k)\bigg] \\ &\quad + \alpha \sum_{i=1}^{q} E\bigg(\sum_{k=1}^{N} H(\mathsf{Q}_i(Y_k^i|Y_1^i, \dots, Y_{k-1}^i)|R_k)\bigg) \\ &\quad + N(c(t, \alpha) + t\varepsilon). \end{aligned} \tag{13}$$

Note that $\mathsf{P} = \prod_{k=1}^N \mathsf{P}(X_k|X_1, \dots, X_{k-1})$, $\mathsf{Q}_i = \prod_{i=1}^N \mathsf{Q}_i(Y_k^i|Y_1^i, \dots, Y_{k-1}^i)$ and let $\mathsf{R} = \prod_{k=1}^N R_k$ be any product measure on Ω^N. Taking $\varepsilon \downarrow 0$, (3) follows from (13) by the well known chain rule for relative entropies $[H(\mathsf{P}|\mathsf{R}) = \sum_{k=1}^N EH(\mathsf{P}(X_k|X_1, \dots, X_{k-1})|R_k)$; cf. [3], Lemma 4.4.7]. □

In particular, corresponding to the functionals d_q, d_h and d_α are the basic "distance" functionals

$$\Delta_q(Q, P) = \int_{\tilde{\Omega}} \left(1 - q\frac{dQ}{dP}\right)_+ dP, \tag{14}$$

$$\Delta_h(Q, P) = \sup\left\{\int \hat{g}\, dP - \int g dQ;\ \hat{g} \in L_1(P),\ g \in L_1(Q), \right. \\ \left. \hat{g}(x) - g(y) \le h(x, y)\ \forall x, y \in \Omega\right\} \tag{15}$$

and

$$\Delta_\alpha(Q, P) = \int_{\tilde{\Omega}} \xi_\alpha\left(\left(1 - \frac{dQ}{dP}\right)_+\right) dP, \tag{16}$$

respectively [where $\tilde{\Omega}$ is such that $P(\tilde{\Omega}^c) = 0$ and dQ/dP exists on $\tilde{\Omega}$].

The next proposition, which is of independent interest, provides the information inequalities of the type (11), relating Δ_q, Δ_h and Δ_α with the relative entropy.

PROPOSITION 1. (i) *For every choice of probability measures P, Q, R on Ω,*

$$t\Delta_\alpha(Q, P) \le H(P|R) + \alpha H(Q|R) \tag{17}$$

provided $t = 1$ and ξ_α satisfies (2).

(ii) *Inequality (17) holds for Δ_q when $q > 1$, $\alpha > 0$ and t is the unique positive solution of*

$$1 + \alpha = e^t + \alpha e^{-qt/\alpha}, \tag{18}$$

whereas for $q = 1$, any $\alpha > 0$, $t > 0$,

$$t\Delta_q(Q, P) \le H(P|R) + \alpha H(Q|R) + \log a(t, \alpha), \tag{19}$$

where $a(t, \alpha)$ is determined as in (4).

(iii) *For h bounded and $b(t, h, R)$ of (5),*

$$t\Delta_h(Q, P) \le H(P|R) + H(Q|R) + b(t, h, R). \tag{20}$$

REMARK 3. Existence and uniqueness of the positive solution t of (18) for $\alpha > 0$ and $q > 1$ is standard [solving $E(\exp(tZ)) = 1$ for bounded random variable Z such that $E(Z) < 0$ and $P(Z > 0) > 0$, taking here $P(Z = 1) = 1 - P(Z = -q/\alpha) = 1/(1+\alpha)$]. Since $\alpha(1 - \exp(-qt/\alpha))$ increases with respect to both q and α, so does the solution t of (18), with $t = \log q$ in case $\alpha = q$.

REMARK 4. Setting in (19), $t = \sqrt{8H(P|R)}$ and $\alpha = \sqrt{H(P|R)/H(Q|R)}$, we recover Pinsker's inequality $\Delta_1(P, Q) \le \sqrt{H(P|R)/2} + \sqrt{H(Q|R)/2}$ by using the bound $\log a(t, \alpha) \le t^2(1 + \alpha^{-1})/8$ of [9], Lemma 2.2.2. Avoiding the latter bound, we can improve on Pinsker's inequality. For example, when $H(P|R) = H(Q|R) = H$ and α and t are as before, (19) reads

$$\Delta_1(Q, P) \le \min\left\{\frac{H + \log\cosh\sqrt{2H}}{\sqrt{2H}},\ 1\right\}.$$

The proof of Proposition 1 relies on the following elementary lemma.

LEMMA 3. (a) *For any probability measures P_0, P_1, R on Ω and $\beta \in [0, 1]$,*

$$\begin{aligned} &\beta H(P_1|R) + (1-\beta)H(P_0|R) \\ &\qquad = \beta H(P_1|P_\beta) + (1-\beta)H(P_0|P_\beta) + H(P_\beta|R), \end{aligned} \tag{21}$$

where $P_\beta = \beta P_1 + (1-\beta)P_0$.

(b) *For any P, Q and $\alpha > 0$*

$$\inf_R\{H(P|R) + \alpha H(Q|R)\} = \int_{\tilde\Omega} \phi_\alpha\left(\frac{dQ}{dP}\right) dP + Q(\tilde\Omega^c)v_\alpha, \tag{22}$$

where $v_\alpha = \alpha\log(1+\alpha^{-1})$ and

$$\phi_\alpha(x) = \alpha x\log x - (1+\alpha x)\log\left(\frac{1+\alpha x}{1+\alpha}\right). \tag{23}$$

PROOF. (a) The cases of $\beta = 0$ and $\beta = 1$ are trivial. When $\beta \in (0, 1)$, unless $P_0 \ll R$ and $P_1 \ll R$, both sides of (21) are infinite. Hence, let $f_i = dP_i/dR$, $i = 0, 1$, and $f_\beta = \beta f_1 + (1-\beta)f_0 = dP_\beta/dR$. Then

$$\begin{aligned} &\beta H(P_1|R) + (1-\beta)H(P_0|R) \\ &\qquad = \int[\beta f_1\log f_1 + (1-\beta)f_0\log f_0]\, dR \\ &\qquad = \int\left[\beta f_1\log\frac{f_1}{f_\beta} + (1-\beta)f_0\log\frac{f_0}{f_\beta} + f_\beta\log f_\beta\right] dR \\ &\qquad = \beta H(P_1|P_\beta) + (1-\beta)H(P_0|P_\beta) + H(P_\beta|R). \end{aligned}$$

(b) Applying (21) for $\beta = 1/(1+\alpha)$, $P_1 = P$, $P_0 = Q$, since $H(P_\beta|R) \ge 0$, it follows that the infimum in the lhs of (22) is obtained for $R_\alpha = (P+\alpha Q)/(1+\alpha)$. With $f = dQ/dP$ on $\tilde\Omega$, noting that

$$H(P|R_\alpha) = \int_{\tilde\Omega}\log\left(\frac{1+\alpha}{1+\alpha f}\right) dP$$

and

$$H(Q|R_\alpha) = \int_{\tilde\Omega} f\log\left(\frac{(1+\alpha)f}{1+\alpha f}\right) dP + Q(\tilde\Omega^c)\log\left(\frac{1+\alpha}{\alpha}\right),$$

it is easy to check that $H(P|R_\alpha) + \alpha H(Q|R_\alpha)$ is identical to the rhs of (22). □

PROOF OF PROPOSITION 1. (i) Without loss of generality we assume equality in (2) for every $u \in [0,1]$. Then, by (16) and (23), for every P, Q in $\tilde{\Omega}$

$$\Delta_\alpha(Q,P) = \int_{\tilde{\Omega}} \phi_\alpha\left(\frac{dQ}{dP}\right) 1_{\{dQ/dP<1\}}\, dP. \tag{24}$$

Since $\phi_\alpha(1) = 0$, $\phi'_\alpha(1) = 0$ and $\phi''_\alpha(x) = \alpha/(x(1+\alpha x))$ is positive for $x \geq 0$, it follows that $\phi_\alpha(x) \geq 0$ for $x \geq 0$. Hence, (17) follows by comparing (22) with (24).

(ii) Let $h_q(x) = \phi_\alpha(x) - t(1-qx)_+$ for $q \geq 1$, $t > 0$ and $\alpha > 0$. Fixing $\alpha > 0$, $q > 1$, (18) has a unique positive solution [since $k(t) = e^t + \alpha e^{-qt/\alpha}$ is convex, $k(0) = (1+\alpha)$, $k'(0) = 1 - q < 0$ and $\lim_{t\to\infty} k(t) = \infty$]. Since $h'_q(x) = \alpha \log((1+\alpha)x/(1+\alpha x)) + tq$ is increasing on $[0, q^{-1}]$, the global minimum of $h_q(\cdot)$ on $[0, q^{-1}]$ is at $x^* = 1/(q \vee ((1+\alpha)e^{qt/\alpha} - \alpha))$. For $x \geq q^{-1}$, $h_q(x) = \phi_\alpha(x)$ is nonnegative by part (i) above, and taking t as determined by (18) we have that $h_q(x^*) = \log(1 + \alpha - \alpha e^{-qt/\alpha}) - t$ is zero. Since $h_q(x)$ is nonnegative for $x \geq 0$, we arrive at (17) by comparing (14) with (22).

By (14) and (22) we also arrive at (19) provided that for any $\alpha > 0$, $t > 0$,

$$-\log a(t,\alpha) = \inf_{\{f\geq 0:\ \int_{\tilde{\Omega}} f dP \leq 1\}} \left\{ \int_{\tilde{\Omega}} h_1(f)\, dP + \left(1 - \int_{\tilde{\Omega}} f\, dP\right) v_\alpha \right\}. \tag{25}$$

Since $h_1(x)$ is convex on $[0,1]$ and also on $[1,\infty)$, the rhs of (25) is minimal when $f = x_1 1_B + x_2 1_{B^c}$ for some $x_2 \geq 1 \geq x_1$ and $p = P(B) \in [0,1]$ such that $x_1 p + x_2(1-p) \leq 1$. For f of this form, the expression in the rhs of (25) is $v_\alpha + p[\phi_\alpha(x_1) - (1 - x_1) - x_1 v_\alpha] + (1-p)[\phi_\alpha(x_2) - x_2 v_\alpha]$ which is monotone decreasing with respect to x_2. Thus, we may set $p = (x_2 - 1)/(x_2 - x_1)$ for which (25) amounts to $-\log a(t,\alpha) = \inf_{x_2 \geq 1 \geq x_1 \geq 0} k(x_1, x_2)$, where

$$k(x_1,x_2) = \left(\frac{x_2 - 1}{x_2 - x_1}\right)\phi_\alpha(x_1) + \left(\frac{1 - x_1}{x_2 - x_1}\right)\phi_\alpha(x_2) - \frac{(x_2-1)(1-x_1)}{(x_2 - x_1)} t. \tag{26}$$

Differentiating k, it is not hard to check that $\nabla k(x_1, x_2) = 0$ at the unique point $x_1^* = \alpha^{-1}(1 - e^{-t})/(e^{t/\alpha} - 1)$, $x_2^* = \alpha^{-1}(e^t - 1)/(1 - e^{-t/\alpha})$ at which the Hessian of k is positive definite (note also that $x_2^* > 1 > x_1^* > 0$). Moreover, $k(x_1^*, x_2^*) = -\log a(t,\alpha)$ and the minimal value of $k(\cdot,\cdot)$ at the boundaries $x_1 = 0$ or $x_2 \to \infty$ exceeds $k(x_1^*, x_2^*)$.

(iii) With h bounded above, $\hat{g}(x) - g(y) \leq th(x,y)$ implies that g is bounded below. Hence, moving a constant from $\hat{g}$ to g, with no loss of generality $g \geq 0$. Suppose $f = dP/dR$ exists and $\hat{g} \in L_1(P)$ is such that $\int e^{\hat{g}}\, dR < \infty$. Define S via $dS/dR = e^{\hat{g}}/(\int e^{\hat{g}}\, dR)$. Then,

$$0 \leq H(P|S) = -\int \hat{g}\, dP + H(P|R) + \log \int e^{\hat{g}}\, dR. \tag{27}$$

Likewise, for any Q, $g \in L_1(Q)$ and any R

$$0 \leq \int g\, dQ + H(Q|R) + \log \int e^{-g}\, dR. \tag{28}$$

Since $t\Delta_h = \Delta_{th}$ for any $t > 0$, adding (27) and (28) we obtain (20). □

With Proposition 1 established, the next proposition is key to the proof of parts (i) and (ii) of Theorem 1 and is of some independent interest.

PROPOSITION 2. (i) *For any* $q \in \mathbb{N}$, *and probability measures* $Q_1, \dots, Q_q, P$,

$$\inf_{\pi \in \mathcal{M}_1(Q_1,\dots,Q_q,P)} \pi(X \notin \{Y^i, i = 1, \dots, q\}) = \Delta_q\left(\frac{1}{q}\sum_{i=1}^{q} Q_i, P\right). \tag{29}$$

(ii) *For any* ξ_α, Q, P,

$$\inf_{\pi \in \mathcal{M}_1(Q,\,P)} \int \xi_\alpha(\pi(Y \neq X | X = x))\, dP(x) \le \Delta_\alpha(Q, P),$$

with equality when ξ_α *is convex and nondecreasing.*

REMARK 5. For $q = 1$, (29) is the classical characterization of the total variation distance.

PROOF OF PROPOSITION 2. (i) Hereafter let $Q_{q+1} = P$ and for nonnegative measures S, T of finite total mass, let $(S - T)_+$ denote the positive part of the signed measure $S - T$, while $S \wedge T$ denotes the nonnegative measure $S - (S - T)_+ = T - (T - S)_+$. For $r = 1, \dots, q + 1$ let $\nu_r = [(P - \sum_{i=1}^{r-1} Q_i)_+ \wedge Q_r](\Omega)$. Since

$$\sum_{r=1}^{q+1}\left(P - \sum_{i=1}^{r-1} Q_i\right)_+ \wedge Q_r = P, \tag{30}$$

in particular, $\sum_{r=1}^{q+1} \nu_r = 1$. Also note that for $r = 1, \dots, q$,

$$\left(P - \sum_{i=1}^{r-1} Q_i\right)_+ \wedge Q_r + \left(\sum_{i=1}^{r} Q_i - P\right)_+ \wedge Q_r = Q_r \tag{31}$$

and, in particular, for $r = 1, \dots, q$,

$$\left[\left(\sum_{i=1}^{r} Q_i - P\right)_+ \wedge Q_r\right](\Omega) = 1 - \nu_r.$$

Suppose $0 < \nu_r < 1$, $r = 1, \dots, q+1$ and that $(\Omega, \mathcal{F})$ is rich enough to support the independent random variables $\{W_r\}_{r=1}^{q+1}$ and $\{Z_r\}_{r=1}^{q}$ with $W_r \sim \nu_r^{-1}(P - \sum_{i=1}^{r-1} Q_i)_+ \wedge Q_r$ and $Z_r \sim (1 - \nu_r)^{-1}(\sum_{i=1}^{r} Q_i - P)_+ \wedge Q_r$. Let $I \in \{1, \dots, q+1\}$ be chosen independently of all these variables according to the probabilities $\{\nu_1, \dots, \nu_{q+1}\}$. Finally let $X = W_I$ and $Y^r = (W_r 1_{I=r} + Z_r 1_{I \neq r})$ for $r = 1, \dots, q$.

The identity (30) implies that $X \sim P$ while (31) results with $Y^r \sim Q_r$. Also note that in this coupling $(Y^1, \dots, Y^q, X) \sim \pi \in \mathcal{M}_1(Q_1, \dots, Q_q, P)$ is such that

$$\pi(X \notin \{Y^i,\ i = 1, \dots, q\}) \le \nu_{q+1} = \left(P - \sum_{i=1}^{q} Q_i\right)_+(\Omega) = \Delta_q\left(q^{-1}\sum_{i=1}^{q} Q_i, P\right).$$

If $\nu_r = 0$, then we do not need the variable W_r for the construction of $Y^1, \dots, Y^q, X$, whereas for $\nu_r = 1$ we never use Z_r. Hence, we have just established the less than or equal to direction in (29).

To show the converse, let $D = \{x \in \tilde{\Omega}\colon \sum_{i=1}^q dQ_i/dP(x) \le 1\}$. Then, for $\pi \in \mathscr{M}_1(Q_1, \dots Q_q, P)$,

$$\pi(X \notin \{Y^i, i = 1, \dots, q\}) \ge \pi(\{X \in D\} \cap_i \{Y^i \notin D\})$$
$$\ge P(D) - \sum_{i=1}^q Q_i(D) = \Delta_q\left(q^{-1} \sum_{i=1}^q Q_i, P\right).$$

(ii) For $q = 1$ the construction of part (i) above results with $\pi(Y^1 \neq X, X \in \cdot) = (P - Q)_+(\cdot)$, implying that $\int \xi_\alpha(\pi(Y \neq X | X = x))\, dP(x) = \Delta_\alpha(Q, P)$. Conversely, fix $\varepsilon > 0$ and let $D_i = \{x \in \tilde{\Omega}\colon 1 - dQ/dP(x) \in [i\varepsilon, (i+1)\varepsilon)\}$, $i = 0, 1, \dots$. Then, for any $\pi \in \mathscr{M}_1(Q, P)$, by convexity and monotonicity of ξ_α,

$$\int \xi_\alpha(\pi(Y \neq X | X = x))\, dP(x) \ge \sum_i P(D_i) \xi_\alpha(1 - Q(D_i)/P(D_i))$$
$$\ge \sum_i \xi_\alpha(i\varepsilon) P(D_i).$$

Taking $\varepsilon \downarrow 0$ the rhs converges to $\Delta_\alpha(Q, P)$. □

The next duality lemma, which is a special case of [8], Theorem 4, is needed for the proof of part (iii) of Theorem 1.

LEMMA 4. *Suppose P, Q are probability measures on a Polish space Ω and $h \in B(\Omega \times \Omega)$. Then*

$$\inf_{\pi \in \mathscr{M}_1(Q,P)} \int h(x, y)\, d\pi(x, y) = \Delta_h(Q, P). \tag{32}$$

PROOF OF THEOREM 1. (i) Without loss of generality assume equality holds in (2) for every $u \in [0, 1]$. Fix $\varepsilon > 0$, P, Q and a product measure R, using hereafter the notation $P_k(\cdot) = \mathsf{P}(X_k \in \cdot | X_1, \dots, X_{k-1})$ and $Q_k(\cdot) = \mathsf{Q}(Y_k \in \cdot | Y_1, \dots, Y_{k-1})$ for $k = 1, \dots, N$. Fix X of law P. By the convexity of ξ_α determined above, and applying part (ii) of Proposition 2 sequentially for (Q_k, P_k), $k = 1, \dots, N$, there is $Y_k \in \sigma(X_k, Y_1, \dots, Y_{k-1}, X_1, \dots, X_{k-1})$ such that the joint law π of (Y, X) is in $\mathscr{M}_N$(Q, P) satisfying

$$E\xi_\alpha(\pi(Y_k \neq X_k | \mathsf{X})) \le E\xi_\alpha(\pi(Y_k \neq X_k | \mathsf{X}, Y_1, \dots, Y_{k-1}))$$
$$= E\xi_\alpha(\pi(Y_k \neq X_k | Y_1, \dots, Y_{k-1}, X_1, \dots, X_k)) \tag{33}$$
$$\le E\Delta_\alpha(Q_k, P_k) + \varepsilon.$$

The proof is completed by combining part (i) of Proposition 1 with Lemma 2 [compare (17) and (33) with (11) and (12), respectively].

(ii) Fix $\varepsilon > 0$, P, $Q_1, \dots, Q_q$ and a product measure R. Now let X have law P and $Y^i = (Y_1^i, \dots, Y_N^i)$ have law Q_i for $i = 1, \dots, q$ with $Q_{k,i}$ denoting the law of Y_k^i conditioned upon $(Y_1^i, \dots, Y_{k-1}^i)$. Applying part (i) of Proposition 2 sequentially for $(Q_{k,1}, \dots, Q_{k,q}, P_k)$, $k = 1, \dots, N$, there exist $(Y_k^1, \dots, Y_k^q, X_k) \in \sigma(Y_1^i, \dots, Y_{k-1}^i, X_1, \dots, X_{k-1})$ such that the joint law π of $(Y^1, \dots, Y^q, X)$ is in $\mathscr{M}_N(Q_1, \dots, Q_q, P)$ satisfying

$$\pi(X_k \notin \{Y_k^i, i = 1, \dots, q\}) \le E\Delta_q\left(q^{-1}\sum_{i=1}^q Q_{k,i}, P_k\right) + \varepsilon. \tag{34}$$

The proof is completed by combining part (ii) of Proposition 1 with Lemma 2 [compare (34) with (12)], noting in the case of $q > 1$ that $H(q^{-1}\sum_{i=1}^q Q_{k,i}|R_k) \le q^{-1}\sum_{i=1}^q H(Q_{k,i}|R_k)$.

(iii) Fix $h \in B(\Omega \times \Omega)$, $\alpha = 1$, $t > 0$, probability measures P, Q, a product measure $R = \prod_k R_k$ and $\varepsilon > 0$. Using the notations of part (i) above, applying Lemma 4 sequentially for (Q_k, P_k), $k = 1, \dots, N$, there is $Y_k \in \sigma(X_k, Y_1, \dots, Y_{k-1}, X_1, \dots, X_{k-1})$ such that the joint law π of (Y, X) is in $\mathscr{M}_N(Q, P)$ satisfying

$$\int h(x_k, y_k)\, d\pi(x, y) \le E\Delta_h(Q_k, P_k) + \varepsilon. \tag{35}$$

The proof is completed by part (iii) of Proposition 1 and Lemma 2 [compare (35) with (12)]. □

3. Extensions of Theorem 1. Extend the "control by q points" by defining for $m = 2, 3, \dots, q$ the decreasing sequence

$$f_{q,m}(A_1, \dots, A_q, x) = \inf_{\{y^i \in A_i\}_{i=1}^q} \sum_{k=1}^N 1_{\{\sum_{i=1}^q 1_{x_k \neq y_k^i} \ge m\}}. \tag{36}$$

Note that $f_{q,q}$ is merely f_q of (2). The corresponding coupling type "distances" between probability measures in Ω^N are then

$$d_{q,m}(Q_1, \dots, Q_q, P) = \inf_{\pi \in \mathscr{M}_N(Q_1, \dots, Q_q, P)} \sum_{k=1}^N \pi\left(\sum_{i=1}^q 1_{X_k \neq Y_k^i} \ge m\right). \tag{37}$$

Extending part (i) of Proposition 2 it can be shown that

$$\begin{aligned} &\inf_{\pi \in \mathscr{M}_1(Q_1, \dots, Q_q, P)} \pi\left(\sum_{i=1}^q 1_{X \neq Y^i} \ge m\right) \\ &\quad = \int_{\tilde{\Omega}} \max_{q \ge p \ge m} \max_{i_1 \neq i_2 \neq \cdots \neq i_p} \left(1 - \frac{1}{p - m + 1}\sum_{j=1}^p \frac{dQ_{i_j}}{dP}\right)_+ dP, \end{aligned} \tag{38}$$

where $P(\tilde{\Omega}) = 1$ and dQ_i/dP exist on $\tilde{\Omega}$ for $i = 1, \dots, q$. With $\Delta_{q,m}(Q_1, \dots, Q_q, P)$ denoting the rhs of (38), extending part (ii) of Proposition 1 it can also be shown that for any probability measures Q_i, P and R on Ω and any $\alpha > 0$,

$q \geq m \geq 2$,

$$t\Delta_{q,m}(Q_1, \ldots, Q_q, P) \leq H(P|R) + \alpha \sum_{i=1}^{q} H(Q_i|R), \tag{39}$$

provided

$$p(1 - \exp(-t\alpha^{-1}/(p - m + 1))) \geq \alpha^{-1}(\exp(t) - 1) \quad \text{for } p = m, \ldots, q. \tag{40}$$

By the same arguments as in the proof of Theorem 1, the inequality (3) holds for $d_{q,m}$, $m = 2, 3, \ldots, q$, $C(t, \alpha) = 0$ and any $t > 0$ satisfying (40). In particular, for $m = q$ we recover part (ii) of Theorem 1. By [10], Lemma 5.6, for $m = 2$ and $\alpha = 1/q$ the condition (40) applies to $t > 0$ such that $e^{tq/2} + e^{-tq} = 2$. Consequently, we recover [10], Theorem 5.4, by using Lemma 1. The resulting concentration inequalities for all other choices of α, m seem to be new.

It may be of independent interest to extend part (i) of Proposition 2 by considering X^j, $j = 1, \ldots, r$, for $r \geq 2$. For example, with $Q = q^{-1} \sum_{i=1}^{q} Q_i$ and $P = r^{-1} \sum_{j=1}^{r} P_j$ it can be shown that

$$\inf_{\pi \in \mathscr{M}_1(Q_1, \ldots, Q_q, P_1, \ldots, P_r)} \pi\big(\{X^j\}_{j=1}^{r} \cap \{Y^i\}_{i=1}^{q} = \varnothing\big) = r[\Delta_{q/r}(Q, P) - (1 - r^{-1})]_+.$$

Acknowledgments. The author thanks Michel Ledoux, Katalin Marton and Michel Talagrand for providing preprints of [4], [6], [7], [9] and [10], and Ofer Zeitouni for the many helpful discussions on the content of this work.

REFERENCES

[1] DEMBO, A. and ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston.

[2] DEMBO, A. and ZEITOUNI, O. (1996). Transportation approach to some concentration inequalities in product spaces. *Elect. Comm. in Probab.* 1 83–90.

[3] DEUSCHEL, J. D. and STROOCK, D. W. (1989). *Large Deviations*. Academic Press, Boston.

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

[5] MARTON, K. (1986). A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* IT-32 445–446.

[6] MARTON, K. (1996). Bounding $\bar{d}$-distance by information divergence: a method to prove measure concentration. *Ann. Probab.* 24 857–866.

[7] MARTON, K. (1996). A measure concentration inequality for contracting Markov chains. *GAFA* 6 556–571.

[8] RAMACHANDAN, D. and RÜSCHENDORF, L. (1995). A general duality theorem for marginal problems. *Probab. Theory Related Fields* 101 311–319.

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

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

[11] TALAGRAND, M. (1996). Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.* 6 587–600.

DEPARTMENTS OF MATHEMATICS AND STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: amir@playfair.stanford.edu