Electron. Commun. Probab. **20** (2015), no. 4, 1–12.
DOI: 10.1214/ECP.v20-3431
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

Optimal transport and Rényi informational divergence

Sergey G. Bobkov[*] Ying Ding[†]

Abstract

Transport-entropy inequalities are considered in terms of Rényi informational divergence.

Keywords: Optimal transport; transport-entropy inequalities; Rényi informational divergence.
AMS MSC 2010: Primary 60E05; 60E15, Secondary 46Gxx.
Submitted to ECP on April 4, 2014, final version accepted on September 7, 2014.

1 Introduction

Let (E, ρ) be a separable metric space. The Kantorovich distance between (Borel) probability measures μ and ν on E is defined by

$$W_1(\mu, \nu) \ = \ \inf_{\pi} \iint \rho(x, y)\, d\pi(x, y)$$

with infimum taken over all measures π on the product space $E \times E$ having μ and ν as marginal projections. One often tries to relate it to more tractable distance-like quantities or measures of deviation such as the Kullback-Leibler informational divergence (relative entropy)

$$D(\nu|\mu) = \int f \log f\, d\mu,$$

assuming that ν is absolutely continuous with respect to μ ($\nu \ll \mu$) and has density $f = \frac{d\nu}{d\mu}$. In particular, relations

$$W_1(\mu, \nu) \leq K \sqrt{D(\nu|\mu)} \tag{1.1}$$

form an important class of transport-entropy inequalities, with interesting connections to high-dimensional phenomena, limit theorems and other problems of Probability, Analysis and Geometry (cf. e.g. [7], [8], [9]). The measure μ in (1.1) is commonly fixed and is called a reference measure, while ν is arbitrary.

The validity of the inequality (1.1) with some (finite) constant K is known to be equivalent to the property

$$\int e^{c\rho(x,x_0)^2}\, d\mu(x) < \infty \tag{1.2}$$

[*]University of Minnesota, Minneapolis, USA. E-mail: bobkov@math.umn.edu
[†]Zhejiang University of Technology, P. R. China. E-mail: yding@zjut.edu.cn

which should hold with some $c > 0$ and $x_0 \in E$ ([1], [3]). This subgaussian condition may occur to be rather restrictive in applications, since for the finiteness of W_1 one only needs the finiteness of the first moment $\int \rho(x, x_0)\, d\mu(x)$. Therefore, it is natural to consider weaker variants of (1.1) with other informational distances so that to involve a larger class of reference distributions μ. As it turns out, to this aim the Rényi divergence power of order $\alpha > 1$,

$$D_\alpha(\nu|\mu) = \frac{1}{\alpha - 1}\left[\int f^\alpha\, d\mu - 1\right],$$

can replace D. It is related to Rényi's entropy like the Kullback-Leibler divergence is related to Shannon's entropy. Note that $0 \le D_\alpha \le \infty$, the function $\alpha \to D_\alpha$ is non-decreasing, and that $\lim_{\alpha \downarrow 1} D_\alpha = D$ (as long as $D_\alpha < \infty$ for some $\alpha > 1$). We refer to [5,6] for an account of basic properties of these functionals.

The aim of this note is to derive the following characterization complementing the equivalence of (1.1) and (1.2).

Theorem 1.1. *Let $1 < \alpha \le 2$. A probability measure μ on E satisfies the relation*

$$W_1(\mu, \nu) \le K\sqrt{D_\alpha(\nu|\mu)} \tag{1.3}$$

with some constant K for all probabilities $\nu \ll \mu$, if and only if, for some $x_0 \in E$,

$$\sup_{r>0}\left[r^{\alpha^*-2}\int_{\rho(x,x_0)\ge r} \rho(x, x_0)^{\alpha^*}\, d\mu(x)\right] < \infty. \tag{1.4}$$

Here and below $\alpha^* = \frac{\alpha}{\alpha-1}$ stands for the conjugate power. Note that $\alpha^* \ge 2$ for $\alpha \in (1, 2]$.

The weakest case in (1.3) is $\alpha = 2$, which is possible according to (1.4), if and only if $\int \rho(x, x_0)^2\, d\mu(x) < \infty$. Thus, Theorem 1.1 involves all probability distributions with finite second moment. More generally, (1.4) is fulfilled as long as

$$\int \rho(x, x_0)^{2\alpha^*-2}\, d\mu(x) < \infty. \tag{1.5}$$

However, this moment condition is strictly stronger than (1.4) in general.

Since (1.4) may not be true for $\alpha > 2$, a different description should appear for (1.3). However, this case turns out to be essentially the same as $\alpha = 2$.

Theorem 1.2. *Let $\alpha \ge 2$. The relation (1.3) holds with some constant K for all probabilities $\nu \ll \mu$, if and only if $\int \rho(x, x_0)^2\, d\mu(x) < \infty$ for some $x_0 \in E$.*

Example 1.3. Let μ be the generalized Cauchy distribution on the Euclidean space $E = \mathbb{R}^n$ (equipped with the Euclidean distance ρ), i.e., with density with respect to Lebesgue measure

$$\frac{d\mu(x)}{dx} = \frac{c}{(1 + |x|^2)^{(n+d)/2}}.$$

Here d is a real parameter (necessarily $d > 0$ for the integrability reason), and c is a normalizing constant. Clearly, μ has finite second moment, when $d > 2$. In this case, (1.4) is telling us that μ satisfies the transport-entropy inequality (1.3), if and only if $\alpha \ge 1 + \frac{2}{d}$. Note that (1.5) excludes the critical value $\alpha = 1 + \frac{2}{d}$.

One should mention that there exist results relating the transport cost $W_1(\mu, \nu)$ to other quantities depending on the distribution under μ of the density f of ν. For example, [2] provides a characterization for the inequalities including

$$W_1(\mu, \nu) \le C\left(\int f^\alpha\, d\mu\right)^{1/\beta} \quad (\beta \ge \alpha \ge 1).$$

Here, the right-hand side has a strong relationship with the Rényi divergence power. However, it does not have the meaning of a distance, and the inequality itself should be viewed from a different point of view.

Let us also comment on the related functional – the Rényi divergence

$$d_\alpha(\nu|\mu) = \frac{1}{\alpha-1}\,\log\int f^\alpha\,d\mu, \qquad \alpha>1.$$

We have $D_\alpha = \frac{1}{\alpha-1}\,(e^{(\alpha-1)d_\alpha}-1)$, so d_α and D_α are equivalent, when these quantities are small. Since in general $D \le d_\alpha \le D_\alpha$, one may wonder whether or not the transport-entropy inequality (1.3) can be replaced with a sronger relation

$$W_1(\mu,\nu) \le K\sqrt{d_\alpha(\nu|\mu)}.$$

However, this inequality turns out to be equivalent to the limit case $\alpha=1$. That is, it holds if and only if the subgaussian integrability condition (1.2) is fulfilled (cf. Remark 4.2 below).

The paper is organized as follows. We start with the study of the Rényi divergence power as a convex functional on the space of densities and provide its description in the form of the supremum of certain linear functionals (Section 2); some immediate consequences are then developed in Section 3. In sections 4-5 we prove Theorems 1.1-1.2, actually in a more quantified form of two-sided bounds on the optimal constant K in (1.3). In particular, for $p\ge 2$, we consider the quantities $K_p = K_p(E,\rho,\mu)\ge 0$ given by

$$K_p^{2p-2} \,=\, \sup_u\,\sup_{r>0}\left[r^{p-2}\int_{|u|\ge r}|u|^p\,d\mu\right], \tag{1.6}$$

where the first supremum is running over all functions u on E with Lipschitz semi-norm $\|u\|_{\rm Lip}\le 1$ and μ-mean zero. It will be shown that $K\sim K_{\alpha^*}$ within factors depending on $\alpha\in(1,2]$, only.

2 Linearization of the Rényi divergence power

Denote by $\mathcal{P}(\mu)$ the collection of all (probability) densities f on an abstract measurable space E with respect to a given probability measure μ. Being convex on $\mathcal{P}(\mu)$, the entropy functional D admits a well-known sup-linear representation, namely

$$D(\nu|\mu) = \int f\,\log f\,d\mu = \sup\left\{\int fg\,d\mu : \int e^g\,d\mu\le 1\right\}.$$

In other words,

$$\int g\,d\nu \le D(\nu|\mu)$$

for all $\nu\ll\mu$, if and only if $\int e^g\,d\mu\le 1$. As a first step towards Theorem 1.1, we derive a similar description for the Rényi divergence power of an arbitrary order $\alpha>1$.

In the sequel, we write $t_+=\max(t,0)$ and denote by $L^p(\mu)$ the usual Lebesgue space of all measurable functions g on E with finite norm $\|g\|_p = (\int|g|^p\,d\mu)^{1/p}$, $p\ge 1$.

Theorem 2.1. *Assume that $g_+\in L^{\alpha^*}(\mu)$. The relation*

$$\int g\,d\nu \;\le\; D_\alpha(\nu|\mu) \tag{2.1}$$

holds for any probability measure $\nu\ll\mu$, if and only if

$$\int(g-c)_+^{\alpha^*}\,d\mu \;\le\; -(\alpha^*)^{\alpha^*}\big(c+\alpha^*-1\big), \tag{2.2}$$

where c is a unique solution to the equation

$$\int (g-c)_+^{\alpha^*-1}\, d\mu \ = \ (\alpha^*)^{\alpha^*-1}. \tag{2.3}$$

As an equivalent description, Theorem 2.1 admits the following analog.

Theorem 2.2. *Assume that $g_+ \in L^{\alpha^*}(\mu)$. The relation (2.1) holds for any probability measure $\nu \ll \mu$, if and only if the condition (2.2) is fulfilled for at least one constant c.*

We split the proof into two steps. On $\mathcal{P}(\mu)$ introduce the concave functional

$$Tf = \int fg\, d\mu - \frac{1}{\alpha-1}\,\Big[\int f^\alpha\, d\mu - 1\Big]$$

with the convention that $Tf = -\infty$ in case $\int f^\alpha\, d\mu = \infty$. Note that Tf is just the difference between the left and right-hand sides of (2.1).

Lemma 2.3. *If $g_+ \in L^{\alpha^*}(\mu)$, the functional T is bounded above on $\mathcal{P}(\mu)$ and attains maximum at some function $f \in \mathcal{P}(\mu) \cap L^\alpha(\mu)$.*

Proof. By Hölder's inequality,

$$|Tf| \ \ \le \ \ \|f\|_\alpha \|g\|_{\alpha^*} - \frac{1}{\alpha-1}\,\Big[\|f\|_\alpha^\alpha - 1\Big] \ \le \ c_0 + c_1 \|g\|_{\alpha^*}^{\alpha^*}$$

up to some constants c_0 and c_1 depending on α, only. Here, when taking the sup over all f, one may assume that $\|f\|_\alpha \le C$ with some large C. Indeed, if $\|f\|_\alpha > C$, the expression

$$\|f\|_\alpha \|g\|_{\alpha^*} - \frac{1}{\alpha-1}\,\Big[\|f\|_\alpha^\alpha - 1\Big]$$

tends to $-\infty$ for $C \to \infty$. Therefore, T is bounded above on $\mathcal{P}$ by the finite constant

$$M \ = \ \sup\Big\{Tf : f \in \mathcal{P}(\mu) \cap L^\alpha(\mu), \ \|f\|_\alpha \le C\Big\}.$$

Take a sequence $f_n \in \mathcal{P}(\mu) \cap L^\alpha(\mu)$ with $\|f_n\|_\alpha \le C$ such that $Tf_n \ \to M$ as $n \to \infty$. The unit ball of L^α is weakly compact, so there is a subsequence $f_{n'}$ weakly convergent to some f with $\|f\|_\alpha \le C$. Necessarily $f \in \mathcal{P}(\mu)$ and

$$\|f\|_\alpha \ \le \ \liminf_{n' \to \infty} \|f_{n'}\|_\alpha.$$

As a result,

$$\begin{aligned}
\limsup_{n' \to \infty} Tf_{n'} \ &= \ \limsup_{n' \to \infty} \Big(\int f_{n'} g\, d\mu - \frac{1}{\alpha-1}\,\Big[\int f_{n'}^\alpha\, d\mu - 1\Big]\Big) \\
&= \ \lim_{n' \to \infty} \int f_{n'} g\, d\mu - \liminf_{n' \to \infty} \frac{1}{\alpha-1}\,\Big[\int f_{n'}^\alpha\, d\mu - 1\Big] \\
&\le \ \int fg\, d\mu - \frac{1}{\alpha-1}\,\Big[\int f^\alpha\, d\mu - 1\Big] \\
&= \ Tf.
\end{aligned}$$

It follows that $Tf = M$. □

Lemma 2.4. *If $g_+ \in L^{\alpha^*}(\mu)$, the maximizer for the functional T is unique and has the form*

$$f \ = \ (\alpha^*)^{1-\alpha^*}\,(g-c)_+^{\alpha^*-1}$$

for some constant c.

Proof. Let f be a maximizer. For $\delta > 0$, put $A_\delta = \{x \in E : f(x) > \delta\}$. Since $f \geq 0$ and $\int f\,d\mu = 1$, we have $\mu(A_\delta) > 0$ for all δ small enough. This will be assumed.

Consider the functions of the form

$$f_\varepsilon = f + \varepsilon u, \quad \varepsilon \in \mathbb{R},$$

where u is a bounded measurable function on E vanishing outside A_δ and such that $\int u\,d\mu = 0$. Then, f_ε will belong to $\mathcal{P}(\mu) \cap L^\alpha(\mu)$ for all sufficiently small ε and hence $Tf_\varepsilon \leq Tf$.

On the other hand, using Taylor's expansion, one can show that

$$\int f_\varepsilon^\alpha\,d\mu = \int f^\alpha\,d\mu + \alpha\varepsilon \int f^{\alpha-1} u\,d\mu + O(\varepsilon^2).$$

Therefore,

$$Tf_\varepsilon - Tf = \varepsilon \int \Big(g - \alpha^* f^{\alpha-1}\Big)\,u\,d\mu + O(\varepsilon^2).$$

Since ε may be both positive and negative (although small), we conclude that

$$\int \Big(g - \alpha^* f_0^{\alpha-1}\Big)\,u\,d\mu = 0$$

for all admissible functions u. But this is only possible when $g - \alpha^* f_0^{\alpha-1} = c$ on A_δ for some constant c. Since $\delta > 0$ may also be arbitrary (although small), this constant c cannot depend on δ. As a result, $g - \alpha^* f^{\alpha-1} = c$ on the set $A_0 = \{x \in E : f(x) > 0\}$. Since $\frac{1}{\alpha-1} = \alpha^* - 1$, the function f has the stated form.

Finally, let us see that the constant c is uniquely determined by the condition $\int f\,d\mu = 1$. Define

$$\varphi(c) \ = \int (g-c)_+^{\alpha^*-1}\,d\mu.$$

This function is continuous and non-increasing on the real line with $\varphi(-\infty) = \infty$, $\varphi(\infty) = 0$. Moreover, it is (strictly) decreasing for

$$c \ \leq \ c_0 = \operatorname{ess\,sup} g,$$

and we have $\varphi(c_0) = 0$. Indeed, if $\varphi(c_1) = \varphi(c_2)$ for $c_1 < c_2$, then $(g - c_1)_+ = (g - c_2)_+$ μ-a.e., that is,

$$\min(g, c_1) = \min(g, c_2) \quad \mu\text{−a.e.}$$

Here the left-hand side is dominated by the right-hand side. But if $g(x) > c_1$ for some $x \in E$, then $\min(g(x), c_1) = c_1$, while $\min(g(x), c_2) > c_1$, so the above equality is impossible. Therefore, necessarily $\mu\{g > c_1\} = 0$ which proves the last assertion. In particular, for any $b > 0$, the equation $\varphi(c) = b$ has a unique solution c. □

Proof of Theorem 2.1. Combining Lemmas 2.3 and 2.4, it remains to look at the value of T on the extreme density $f = (\alpha^*)^{1-\alpha^*}\,(g - c)_+^{\alpha^*-1}$. First, using the property $\int f\,d\mu = 1$, we have

$$\begin{aligned}\int fg\,d\mu \ &= \ (\alpha^*)^{-(\alpha^*-1)} \int (g-c)_+^{\alpha^*-1}\Big((g-c)_+ + c\Big)\,d\mu \\ &= \ (\alpha^*)^{-(\alpha^*-1)} \int (g-c)_+^{\alpha^*}\,d\mu \ + \ c.\end{aligned}$$

Secondly,

$$\begin{aligned}\frac{1}{\alpha-1}\left[\int f^\alpha\,d\mu - 1\right] \ &= \ \frac{(\alpha^*)^{-\alpha^*}}{\alpha-1} \int (g-c)_+^{\alpha^*}\,d\mu \ - \ \frac{1}{\alpha-1} \\ &= \ (\alpha^*)^{-\alpha^*}\,(\alpha^*-1) \int (g-c)_+^{\alpha^*}\,d\mu \ - \ (\alpha^*-1).\end{aligned}$$

Hence,

$$Tf \,=\, (\alpha^*)^{-\alpha^*} \int (g-c)_+^{\alpha^*}\, d\mu \,+\, c + \alpha^* - 1.$$

Using this extreme function (maximizer), one may rewrite the property "$Tf \leq 0$ for all f", that is, (2.1), in terms of D_α, as indicated in (2.2)-(2.3). □

Proof of Theorem 2.2. The function

$$\psi(c) = \int (g-c)_+^{\alpha^*}\, d\mu \,+\, (\alpha^*)^{\alpha^*} c$$

is strictly convex and is differentiable on $\mathbb{R}$, with $\psi(-\infty) = \psi(\infty) = \infty$. It attains minimum at a unique point c, namely – at which

$$\psi'(c) = \alpha^* \bigg[- \int (g-c)_+^{\alpha^*-1}\, d\mu \,+\, (\alpha^*)^{\alpha^*-1} \bigg] = 0.$$

But this is exactly the equation (2.3), while the inequality $\psi(c) \leq -(\alpha^*)^{\alpha^*}(\alpha^*-1)$ being stated at this point coincides with the condition (2.2). □

3 Necessary and sufficient conditions

Since the description given in Theorem 2.1 for the property

$$\int g\, d\nu \;\leq\; D_\alpha(\nu|\mu) \quad \text{for any } \nu \ll \mu \tag{3.1}$$

is somewhat implicit, it would be interesting to get more tractable conditions, necessary and sufficient, even if not simultaneously. Here we mention some of such conditions, together with lower and upper bounds on the constant c appearing in (2.2)-(2.3). To avoid situations when $D_\alpha(\nu|\mu)$ is finite, but the integral in (3.1) does not exist, we assume that $g_+ \in L^{\alpha^*}(\mu)$.

In particular, applying (3.1) to the measure $\nu = \mu$, we get $\int g\, d\mu \leq 0$. A different choice leads to stronger necessary condition

$$\int \Big(1 + \frac{1}{\alpha^*-1}\, g\Big)_+^{\alpha^*-1} d\mu \leq 1. \tag{3.2}$$

On the other hand, choosing $c = -\alpha^*$ in Theorem 2.2, we arrive at the sufficient condition

$$\int \Big(1 + \frac{1}{\alpha^*}\, g\Big)_+^{\alpha^*} d\mu \leq 1. \tag{3.3}$$

As $\alpha \downarrow 1$, both (3.2) and (3.3) are asymptotically optimal. Indeed, in the limit they yield $\int e^g\, d\mu \leq 1$ which is necessary and sufficient for the relation $\int g\, d\nu \leq D(\nu|\mu)$. Nevertheless, being quite explicit and working, (3.2)-(3.3) are not sharp enough to reach simulteneously necessary and sufficient conditions as in Theorem 1.1.

Let us now return to Theorem 2.1 and recall the condition

$$\int (g-c)_+^{\alpha^*}\, d\mu \;\leq\; -(\alpha^*)^{\alpha^*}\big(c + \alpha^* - 1\big), \tag{3.4}$$

where c solves the equation

$$\int (g-c)_+^{\alpha^*-1}\, d\mu \;=\; (\alpha^*)^{\alpha^*-1}. \tag{3.5}$$

Proposition 3.1. *Under* $(3.4)-(3.5)$*, necessarily* $c \leq -\alpha^*$*. Furthermore, if* $g \in L^1(\mu)$*,*

$$c \geq -\alpha^* + \int g\,d\mu \qquad in\ case\ 1 < \alpha \leq 2,$$
$$c \geq -4 + \alpha \int g\,d\mu \qquad in\ case\ \alpha \geq 2.$$

In particular, in the corresponding cases,

$$\int g_+^{\alpha^*}\,d\mu \ \leq \ (\alpha^*)^{\alpha^*}\Big(1 - \int g\,d\mu\Big), \tag{3.6}$$

$$\int g_+^{\alpha^*}\,d\mu \ \leq \ (\alpha^*)^{\alpha^*}\Big(4 - \alpha \int g\,d\mu\Big). \tag{3.7}$$

Proof. The weaker upper bound $c \leq -(\alpha^* - 1)$ immediately follows from (3.4). To refine it, we use (3.5) and apply Markov's inequality to get

$$\begin{aligned}(\alpha^*)^{\alpha^*} \ &= \ \Big[\int (g-c)_+^{1/(\alpha-1)}\,d\mu\Big]^{\alpha} \\ &\leq \ \int (g-c)_+^{\alpha/(\alpha-1)}\,d\mu \ \leq \ -(\alpha^*)^{\alpha^*}\big(c + \alpha^* - 1\big).\end{aligned}$$

Hence, $1 \leq -\big(c + \alpha^* - 1\big)$ proving the first statement.

For the lower bound on c in case $1 < \alpha \leq 2$, using convexity of the function $t \rightarrow (t-c)_+^{\alpha^*-1}$, one can apply Jensen's inequality in (3.5) to get

$$(\alpha^*)^{\alpha^*-1} \ = \ \int (g-c)_+^{\alpha^*-1}\,d\mu \ \geq \ \Big[\int g\,d\mu - c\Big]_+^{\alpha^*-1}.$$

Hence

$$\alpha^* \ \geq \ \Big[\int g\,d\mu - c\Big]_+ \ \geq \ \int g\,d\mu - c$$

which is the first lower bound. Using it in (3.4), we conclude that

$$\begin{aligned}\int g_+^{\alpha^*}\,d\mu \ &\leq \ \int (g-c)_+^{\alpha^*}\,d\mu \\ &\leq \ (\alpha^*)^{\alpha^*}\big(-c - \alpha^* + 1\big) \ \leq \ (\alpha^*)^{\alpha^*}\Big(1 - \int g\,d\mu\Big)\end{aligned}$$

which is (3.6).

In case $\alpha \geq 2$, we start with (3.4) and first simplify it to $\int (g-c)_+^{\alpha^*}\,d\mu \leq (\alpha^*)^{\alpha^*}(-c)$. By Jensen's inequality,

$$\int (g-c)_+^{\alpha^*}\,d\mu \ \geq \ \Big[\int g\,d\mu - c\Big]_+^{\alpha^*}$$

giving

$$\int g\,d\mu - c \ \leq \ \Big[\int g\,d\mu - c\Big]_+ \ \leq \ \alpha^*(-c)^{1/\alpha^*}.$$

Equivalently, substituting $t = -c$, $p = \alpha^*$, $q = \alpha$, $a = -\int g\,d\mu$, we arrive at the relation

$$\varphi(t) \ \equiv \ t - pt^{1/p} \ \leq \ a.$$

This function is convex in $t \geq 0$ and positive for $t > t_0 = p^q$, with $\varphi(t_0) = 0$, $\varphi'(t_0) = 1 - t_0^{-1/q} = 1/q$. Hence, for all $t \geq t_0$,

$$\varphi(t) \geq \varphi(t_0) + \varphi'(t_0)(t - t_0) = \frac{1}{q}\,(t - t_0).$$

Once $\varphi(t) \leq a$ and $t \geq t_0$, we then get $t \leq t_0 + qa$. But $t_0 \leq 4$ whenever $q \geq 2$. Indeed, for the function $\psi(q) = \log t_0 = q \log p$ we have $\psi''(q) = \frac{1}{q(q-1)^2} > 0$, so it is convex. In addition, $\psi'(\infty) = 0$, so it is decreasing. Hence, $\psi(q) \leq \psi(2)$ for all $q \geq 2$, i.e., $p^q \leq 4$. This gives the required upper bound on c. Again, using it in (3.4), we conclude that

$$\begin{aligned} \int g_+^{\alpha^*}\,d\mu &\leq (\alpha^*)^{\alpha^*}\big(-c-\alpha^*+1\big) \\ &\leq (\alpha^*)^{\alpha^*}\Big(5-\alpha^*-\alpha\int g\,d\mu\Big) \ \leq \ (\alpha^*)^{\alpha^*}\Big(4-\alpha\int g\,d\mu\Big) \end{aligned}$$

which is (3.7). In fact, this last bound will not be needed for the proof of Theorem 1.2. □

4 Finiteness of the second moment

We are prepared to turn to Theorems 1.1-1.2 which will be established in a more quantitative form involving the quantities K_p introduced in (1.6). In particular,

$$K_2^2 \,=\, \sup_{u\in\mathcal{L}} \int u^2\,d\mu,$$

where the supremum is taken over the familiy $\mathcal{L}$ of all functions u on E with Lipschitz semi-norm $\|u\|_{\rm Lip} \leq 1$, having μ-mean zero. This quantity is finite, if and only if μ has a finite second moment. Indeed, for the finiteness, it is enough to consider the Lipschitz function $u(x) = \rho(x,x_0) - \int \rho(x,x_0)\,d\mu(x)$ in the definition of K_2 with an arbitrary fixed point $x_0 \in E$.

Recall that we consider the transport-entropy inequality

$$W_1(\mu,\nu) \leq K\sqrt{D_\alpha(\nu|\mu)} \tag{4.1}$$

with an arbitrary probability measure $\nu \ll \mu$. For example, if $\nu = \mu_A$ has a constant density $f = \frac{1}{\mu(A)}\,1_A$, then (4.1) becomes

$$W_1^2(\mu,\mu_A) \leq K^2\,\frac{\mu(A)^{1-\alpha}-1}{\alpha-1}.$$

Taking for A a ball of a sufficiently large radius so that $\mu(A) > 0$, we get $W_1(\mu,\mu_A) < \infty$, while μ_A has finite first moment. Hence, for (4.1) to hold, necessarily the reference measure μ must have a finite first moment. In that case, by a simple approximation argument, there will be no loss of generality to assume in (4.1) that ν have finite first moments, as well.

Theorem 4.1. *Let $\alpha > 1$. Under (4.1), we have $K_2 \leq K\sqrt{\frac{\alpha}{2}}$.*

Proof. By the Kantorovich-Ribinstein theorem, if μ and ν have finite first moments, there is the representation

$$W_1(\mu,\nu) = \sup_u \bigg|\int u\,d\mu - \int u\,d\nu\bigg|,$$

where the supremum is running over all u on E with $\|u\|_{\rm Lip} \leq 1$ (cf. e.g. [4], p.330). Then, (4.1) may equivalently be rewritten as

$$\sup_{u\in\mathcal{L}} \int u\,d\nu \leq K\sqrt{D_\alpha(\nu|\mu)}. \tag{4.2}$$

Given a bounded function h on E such that $\int h\,d\mu = 0$ and $\varepsilon > 0$ small enough, the function $f_\varepsilon = 1 + \varepsilon h$ represents the density of a probability measure, say $\nu = \nu_\varepsilon$, with respect to μ. In this case, (4.2) becomes

$$\varepsilon \sup_{u\in\mathcal{L}} \int uh\,d\mu \leq K\sqrt{D_\alpha(\nu_\varepsilon|\mu)}. \tag{4.3}$$

Furthermore, by Taylor's expansion over ε,

$$D_\alpha(\nu_\varepsilon|\mu) = \frac{1}{\alpha-1}\left[\int f_\varepsilon^\alpha\,d\mu - 1\right] = \frac{\alpha\varepsilon^2}{2}\int h^2\,d\mu + O(\varepsilon^3).$$

Inserting this in (4.3) and letting $\varepsilon \to 0$, we arrive at

$$\int uh\,d\mu \le K\sqrt{\frac{\alpha}{2}}\,\|h\|_2$$

holding for any $u \in \mathcal{L}$. But this is equivalent to $\|u\|_2 \le K\sqrt{\frac{\alpha}{2}}$. □

Remark 4.2. Let us look at the possible sharpening of (4.1) in terms of the Rényi divergence, namely

$$W_1(\mu,\nu) \le K\sqrt{d_\alpha(\nu|\mu)}. \tag{4.4}$$

By the definition, if $\nu = \mu_A$ has a constant density $f = \frac{1}{\mu(A)}\,1_A$ (as before), then

$$d_\alpha(\mu_A|\mu) = \log\frac{1}{\mu(A)}$$

which is independent of α. On the other hand (following Marton's argument), given two measurable sets $A, B \subset E$ at distance $r = \rho(A,B)$, we have $W_1(\mu_A,\mu_B) \ge r$. Applying the triangle inequality for the metric W_1, (4.4) therefore yields

$$r \,\le\, W_1(\mu,\mu_A) + W_1(\mu,\mu_B) \,\le\, K\left(\sqrt{\log\frac{1}{\mu(A)}} + \sqrt{\log\frac{1}{\mu(B)}}\right).$$

In particular, if $\mu(A) \ge \frac{1}{2}$, writing $B = E\backslash A^r$ in terms of the r-neighbourhood A^r of A for the metric ρ, we get

$$1 - \mu(A^r) \,\le\, \frac{1}{2}\,e^{-r^2/(2K^2)}, \qquad r > 0.$$

But this property is equivalent to the subgaussian condition (1.2). Therefore, (4.4) is equivalent to the standard transport-entropy inequality (1.1), corresponding to the order $\alpha = 1$.

5 Theorems 1.1-1.2 and their refinements

Theorem 1.1 allows the following refinement in terms of the quantity

$$K_p^{2p-2} \,=\, \sup_{u\in\mathcal{L}}\,\sup_{r>0}\left[r^{p-2}\int_{|u|\ge r}|u|^p\,d\mu\right], \qquad p \ge 2.$$

Theorem 5.1. *Let $1 < \alpha \le 2$. The best value of K in the transport-entropy inequality*

$$W_1(\mu,\nu) \le K\sqrt{D_\alpha(\nu|\mu)} \tag{5.1}$$

satisfies

$$c_\alpha K_{\alpha^*} \le K \le C_\alpha K_{\alpha^*} \tag{5.2}$$

up to some positive constants c_α and C_α depending on α, only.

We also have $K \sim K_2$ for $\alpha > 2$, up to α-depending factors.

Corollary 5.2. *For $\alpha \ge 2$, the best value of K in (5.1) satisfies $\sqrt{\frac{2}{\alpha}}\,K_2 \le K \le CK_2$ with some absolute constant C.*

Indeed, by (5.1)-(5.2) with $\alpha = 2$, and using the monotonicity of the divergence power with respect to α, we get

$$W_1(\mu,\nu) \leq K\sqrt{D_2(\nu|\mu)} \leq C_2 K_2 \sqrt{D_\alpha(\nu|\mu)}.$$

This gives an upper bound $K \leq C_2 K_2$, while the lower bound is provided by Theorem 4.1.

Note that Theorems 1.1-1.2 are immediately obtained from Theorem 5.1 and Corollary 5.2, since K_{α^*} is finite (with $\alpha \leq 2$), if and only if the expression in (1.4) is finite.

Before turning to the proof of Theorem 5.1, first let us explain how we will connect (5.1) to the relations as in Theorem 2.1, i.e.,

$$\int g\,d\nu \ \leq\ D_\alpha(\nu|\mu). \tag{5.3}$$

As was already mentioned in the previous section, (5.1) may equivalently be rewritten as

$$\Big|\int u\,d\nu\Big| \leq K\sqrt{D_\alpha(\nu|\mu)} \qquad \text{for all } u \in \mathcal{L}.$$

Squaring and using $\sup_\lambda\,(\lambda a - \lambda^2) = \frac{a^2}{4}$ together with the property that $-u \in \mathcal{L}$ for all $u \in \mathcal{L}$, we are reduced to the inequality of the form (5.3). That is, we obtain:

Lemma 5.3. *Let K be a positive constant. If μ and ν have finite first moments and $\nu \ll \mu$, (5.1) is equivalent to the the relation*

$$\int \Big(\frac{2}{K}\,\lambda u - \lambda^2\Big)\,d\nu \ \leq\ D_\alpha(\nu|\mu) \tag{5.4}$$

with arbitrary $u \in \mathcal{L}$ and $\lambda > 0$.

Proof of Theorem 5.1 (lower bound on K). First assume that $u \in L^{\alpha^*}(\mu)$. By Proposition 3.1 with $g = \frac{2}{K}\,\lambda u - \lambda^2$ ($\lambda > 0$), we get (3.6) as a necessary condition for (5.4), namely

$$\int \Big(\frac{2}{K}\,\lambda u - \lambda^2\Big)_+^{\alpha^*}\,d\mu \ \leq\ (\alpha^*)^{\alpha^*}(1+\lambda^2).$$

Restricting the integral to the set $u \geq K\lambda$, so that $\frac{2}{K}\,\lambda u - \lambda^2 \geq \frac{\lambda}{K}\,u$, the above yields

$$\frac{\lambda^{\alpha^*}}{K^{\alpha^*}}\int_{u\geq K\lambda} u^{\alpha^*}\,d\mu \ \leq\ (\alpha^*)^{\alpha^*}(1+\lambda^2).$$

To simplify, assume that $\lambda \geq 1$, in which case we thus get

$$\lambda^{\alpha^*-2}\int_{u\geq K\lambda} u^{\alpha^*}\,d\mu \ \leq\ 2\,(K\alpha^*)^{\alpha^*}.$$

Substituting $\lambda = r/K$ and applying the same inequality to $-u$, we arrive at

$$r^{\alpha^*-2}\int_{|u|\geq r} |u|^{\alpha^*}\,d\mu \ \leq\ 4\,(\alpha^*)^{\alpha^*}K^{2\alpha^*-2}, \qquad r \geq K. \tag{5.5}$$

In case $0 \leq r \leq K$, there is a similar obvious bound

$$r^{\alpha^*-2}\int_{r\leq|u|\leq K} |u|^{\alpha^*}\,d\mu \ \leq\ K^{2\alpha^*-2}.$$

On this step, the assumption $u \in L^{\alpha^*}(\mu)$ can easily be removed: (5.5) can always be applied to centered truncated Lipschitz functions $u_n = v_n - \int v_n\,d\mu$, where $v = u$ in case $|u| \leq n$, and $v = \pm n$ depending on whether $u > n$ or $u < -n$. Letting $n \to \infty$, we arrive at

$$\sup_{r>0} \left[r^{\alpha^*-2} \int_{|u|\geq r} |u|^{\alpha^*}\,d\mu \right] \,\leq\, \big(1 + 4\,(\alpha^*)^{\alpha^*}\big)\,K^{2\alpha^*-2}$$

which yields the left inequality in (5.2) with $c_\alpha^{-1} = 1 + 4\,(\alpha^*)^{\alpha^*}$.

Upper bound on K. By Theorem 2.2 applied with the same function $g = \frac{2}{K}\,\lambda u - \lambda^2$, $\lambda > 0$, we know that the relation (5.4) holds true for all $\nu \ll \mu$, if and only if, for some constant c,

$$\int \Big(\frac{2}{K}\,\lambda u - \lambda^2 - c\Big)_+^{\alpha^*} d\mu \,\leq\, (\alpha^*)^{\alpha^*}(1 - \alpha^* - c). \tag{5.6}$$

Case $0 \leq \lambda \leq 2\sqrt{\alpha^*}$. It will be sufficient to establish the latter with $c = -\alpha^* - \lambda^2$, when (5.6) becomes

$$\xi(\varepsilon) \equiv \int \big(1 + \varepsilon\lambda u\big)_+^{\alpha^*} d\mu \,\leq\, 1 + \lambda^2, \qquad \varepsilon = \frac{2}{K\alpha^*}. \tag{5.7}$$

To obtain it with some $\varepsilon = \varepsilon(K)$ independent of λ and $u \in \mathcal{L}$, we use the definition of $R = K_{\alpha^*}$ to write

$$\int |u|^{\alpha^*}\,d\mu \,\leq\, r^{\alpha^*} + \int_{|u|\geq r} |u|^{\alpha^*}\,d\mu \,\leq\, r^{\alpha^*} + \frac{R^{2\alpha^*-2}}{r^{\alpha^*-2}} \qquad (r > 0).$$

Choosing $r = R$, we get an upper bound $\int |u|^{\alpha^*}\,d\mu \,\leq\, 2R^{\alpha^*}$. In particular,

$$\int u^2\,d\mu \,\leq\, 2R^2. \tag{5.8}$$

Note that the function ξ is convex in the variable ε and is twice continuously differentiable. We have

$$\xi'(\varepsilon) = \alpha^*\lambda \int u\big(1 + \varepsilon\lambda u\big)_+^{\alpha^*-1}\,d\mu,$$

so $\xi(0) = 1$, $\xi'(0) = 0$. In addition,

$$\xi''(\varepsilon) = \alpha^*(\alpha^* - 1)\lambda^2 \int u^2\big(1 + \varepsilon\lambda u\big)_+^{\alpha^*-2}\,d\mu. \tag{5.9}$$

Using $(a + b)^p \leq 2^p\,(a^p + b^p)$ $(a, b, p \geq 0)$ and the assumption on the range of λ, we have a pointwise bound

$$\big(1 + \varepsilon\lambda u\big)_+^{\alpha^*-2} \,\leq\, \big(4\sqrt{\alpha^*}\big)^{\alpha^*-2}\,\Big(1 + \varepsilon^{\alpha^*-2}|u|^{\alpha^*-2}\Big).$$

It then follows from (5.8)-(5.9) that

$$\begin{aligned} \xi''(\varepsilon) \;&\leq\; \alpha^*(\alpha^* - 1)\,\big(4\sqrt{\alpha^*}\big)^{\alpha^*-2}\,\lambda^2 \int \Big(u^2 + \varepsilon^{\alpha^*-2}|u|^{\alpha^*}\Big)\,d\mu \\ &\leq\; 2\alpha^*(\alpha^* - 1)\,\big(4\sqrt{\alpha^*}\big)^{\alpha^*-2}\,\lambda^2\big(R^2 + \varepsilon^{\alpha^*-2}R^{\alpha^*}\big). \end{aligned}$$

By Taylor's expansion for $\xi(\varepsilon)$ up to ε^2, we get that, for all $\varepsilon > 0$,

$$\xi(\varepsilon) \,\leq\, 1 + \alpha^*(\alpha^* - 1)\,\big(4\sqrt{\alpha^*}\big)^{\alpha^*-2}\,\lambda^2\big((\varepsilon R)^2 + (\varepsilon R)^{\alpha^*}\big).$$

It suffices to choose here

$$\varepsilon = \frac{1}{2R\sqrt{\alpha^*(\alpha^*-1)\left(4\sqrt{\alpha^*}\right)^{\alpha^*-2}}}$$

to get that $\xi(\varepsilon) \le 1 + \lambda^2$. This means that (5.7) and thus (5.4) are fulfilled with

$$K = \frac{2}{\alpha\varepsilon} = C_\alpha R, \qquad C_\alpha = \frac{4\sqrt{\alpha^*(\alpha^*-1)\left(4\sqrt{\alpha^*}\right)^{\alpha^*-2}}}{\alpha}. \tag{5.10}$$

Case $\lambda \ge 2\sqrt{\alpha^*}$. We choose in (5.6) the value $c = -\alpha^* - \frac{\lambda^2}{2}$ and then need to show that

$$\eta(\varepsilon) \equiv \int \left(1 + \varepsilon\lambda u - \frac{\lambda^2}{2\alpha^*}\right)_+^{\alpha^*} d\mu \,\le\, 1 + \frac{\lambda^2}{2}, \qquad \varepsilon = \frac{2}{K\alpha^*}. \tag{5.11}$$

Since $\frac{\lambda^2}{2\alpha^*} \ge 2$, the property $1 + \varepsilon\lambda u - \frac{\lambda^2}{2\alpha^*} \ge 0$ is implied by $u \ge \frac{\lambda}{\varepsilon}$. Hence, using the definition of R, for all $\varepsilon > 0$,

$$\eta(\varepsilon) \,\le\, \int (\varepsilon\lambda u)^{\alpha^*}\, 1_{\{u \ge \frac{\lambda}{\varepsilon}\}}\, d\mu \,\le\, (\varepsilon\lambda)^{\alpha^*} \cdot \frac{R^{2\alpha^*-2}}{(\frac{\lambda}{\varepsilon})^{\alpha^*-2}} \,=\, \lambda^2\,(\varepsilon R)^{2\alpha^*-2}.$$

It suffices to choose here $\varepsilon = \frac{1}{2R}$ to get $\xi(\varepsilon) \le \frac{\lambda^2}{2}$. This means that (5.11) and thus (5.4) are fulfilled with

$$K = \frac{2}{\alpha\varepsilon} = \frac{4}{\alpha}\,R.$$

Comparing the two cases, we arrive at the right inequality in (5.2) with constant C_α described in (5.10). □

References

[1] S. G. Bobkov and F. Götze: Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.* **163(1)**, (1999), 1–28. MR-1682772

[2] Y. Ding and X. Zhang: A new kind of modified transportation cost inequalities and polynomial concentration inequalities. *Statist. Probab. Lett.* **81(10)**, (2011), 1524–1534. MR-2818664

[3] H. Djellout, A. Guillin and L. Wu: Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.* **32(3B)**, (2004), 2702–2732. MR-2078555

[4] R. M. Dudley: Real analysis and probability. The Wadsworth & Brooks/Cole Mathematics Series. *Pacific Grove*, CA, 1989, xii+436 pp. MR-0982264

[5] T. van Erven and P. Harremoës: Rényi divergence and majorization. *IEEE International Symposium on Information Theory (ISIT)*, 2010.

[6] T. van Erven and P. Harremoës: Rényi divergence and Kullback-Leibler divergence. arXiv:1206.2459v1. To appear in: *IEEE Transactions on Information Theory*. MR-3225930

[7] N. Gozlan: Integral criteria for transportation-cost inequalities. *Electron. Comm. Probab.* **11**, (2006), 64–77 (electronic). MR-2231734

[8] N. Gozlan and C. Léonard: Transport inequalities. A survey. *Markov Process. Related Fields.* **16(4)**, (2010), 635–736. MR-2895086

[9] C. Villani: Topics in Optimal Transportation. *Amer. Math. Soc.*, Providence, RI, 2003. MR-1964483

Acknowledgments. Research of SB was partially supported by NSF grant DMS-1106530. Research of YD was supported by Department of Education of Zhejiang Province Grant (Y201329636) and Zhejiang University of Technology Grant (Zhaohui Plan).