Communications in **M**athematical **A**nalysis
Volume 14, Number 1, pp. 1–14 (2013) www.math-res-pub.org/cma
ISSN 1938-9787

An Explicit Bound on the Transportation Cost Distance

Aryeh Kontorovich*
Department of Computer Science
Ben-Gurion University
Beer Sheva, Israel 84105

(Communicated by Simeon Reich)

Abstract

We give what appears to be the first explicit, easily computable bound on the transportation cost distance with respect to the weighted Hamming metric. The bound follows from Kantorovich duality and a novel inequality, which amounts to bounding the maximal value of certain linear programs and may be of independent interest. We give two application to concentration of measure for dependent processes and pose some open problems and directions for future work.

AMS Subject Classification: 46B99, 60G42

Keywords: optimal transport, concentration of measure, linear program

1 Introduction

1.1 Background

In 1781, Gaspard Monge considered the following problem: soil is extracted from a number of sites and is to be transported to various construction locations. Assuming a fixed cost for transporting a unit of soil over a unit distance, the objective is to come up with a strategy that minimizes the total cost of the operation.

This problem may be formalized as follows. First, let us normalize the total mass of excavated soil to be one; thus, instead of transporting *pounds* of soil from excavation site x to construction site y we may think in terms of percentages. Let $(\mathcal{X},\rho)$ be a metric space endowed with probability distributions μ and ν. We can use μ to model the distribution of extracted soil at the excavation sites and ν to model the distribution of requisite soil at the construction sites. The distance ρ plays the role of price: it costs $\rho(x,y)$ to move a unit of soil from location x to location y.

*E-mail address: karyeh@cs.bgu.ac.il

A transportation strategy is formalized by the notion of a *coupling*. A coupling of μ and ν is defined to be any distribution π on $\mathcal{X}\times\mathcal{X}$ with marginals μ and ν, respectively:

$$\mu(\cdot) = \int_{\mathcal{X}} \pi(\cdot, \mathrm{d}y) \qquad \text{and} \qquad \nu(\cdot) = \int_{\mathcal{X}} \pi(\mathrm{d}x, \cdot).$$

Intuitively, $\pi(\cdot\,|\,x) \equiv \pi(x,\cdot)/\mu(x)$ is a distribution on $\mathcal{X}$ which corresponds to a prescription of how to divide the unit mass at x among the different locations. The trivial coupling $(\mu\otimes\nu)(x,y) = \mu(x)\nu(y)$ always allocates the unit weight at x according to ν — independently of the location x. In general, many other couplings will exist, and we will denote their collection by $\Pi(\mu,\nu)$.

The cost associated with the coupling π is

$$\int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)\pi(\mathrm{d}x,\mathrm{d}y),$$

and the optimal transportation cost is the one realized by the most parsimonious coupling:

$$T_\rho(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)\pi(\mathrm{d}x,\mathrm{d}y). \tag{1.1}$$

The functional T_ρ is easily verified to be a metric on the space of distributions on $\mathcal{X}$. It is known by a host of names, including transportation or earthmover distance, and occasionally also bears the names of its various discoverers: Kantorovich, Monge, Rubinstein, Wasserstein.

Villani [21, 22] provides a fascinating account of the several independent discoveries of optimal transport and is an encyclopedic source on the subject. The other "founding father" of optimal transport was Leonid Kantorovich, who provided a powerful dual characterization of the transportation distance. Under suitable conditions [22, Theorem 5.10], Kantorovich showed that

$$T_\rho(\mu,\nu) \quad = \quad \sup_{\|f\|_{\mathrm{Lip}}\leq 1} \left(\int_{\mathcal{X}} f\mathrm{d}\mu - \int_{\mathcal{X}} f\mathrm{d}\nu \right), \tag{1.2}$$

where $\|\cdot\|_{\mathrm{Lip}}$ is the Lipschitz semi-norm of $f : \mathcal{X} \to \mathbb{R}$ with respect to ρ.

The significance of optimal transport extends far beyond optimal planning, spanning such diverse fields as probability theory [13], statistics [20], computer science [1, 3, 6, 18, 7, 16] and analysis and PDEs [21].

1.2 Statement of results and related work

We prove an explicit, analytically computable estimate on T_ρ for the case where ρ is an ℓ_1 sum of discrete metrics:

$$\rho(x,y) = \sum_{i=1}^{n} w_i \mathbf{1}_{\{x_i\neq y_i\}}, \qquad x,y\in\Omega^n; \tag{1.3}$$

the latter is also known as a "weighted Hamming" metric. This estimate, stated in Corollary 3.4, upper-bounds $T_\rho(\mu,\nu)$ by $\Psi(\mu-\nu)$, where the Ψ functional is given by a simple closed-form expression (2.1).

Our transport inequality is a consequence of a novel linear programming inequality proved in Theorem 2.2. Another simple consequence of the latter is a measure concentration inequality (Corollary 3.2).

A special case of Theorem 2.2 (with $w_i \equiv 1$) was proved in [9]; however, we were unable to extend that proof to general w, which motivated the approach in this paper. The result we obtain here is both much simpler and significantly more general.

This paper is organized as follows. The linear programming inequality is proved in Section 2. In Section 3.1 we give an application of Theorem 2.2 to the Azuma-Hoeffding method of bounded martingale differences to obtain a concentration inequality that is sensitive to the Hamming weights w. In Section 3.2 we propose a way of combining our inequality with Marton's transportation technique to obtain other concentration inequalities.

1.3 Notation

Throughout this paper, Ω will denote a finite set. Random variables are capitalized (X), specified sequences (words) are written in lowercase ($x \in \Omega^n$), the shorthand $X_i^j = (X_i, \ldots, X_j)$ is used for all sequences, and word concatenation is denoted using the multiplicative notation: $x_i^j x_{j+1}^k = x_i^k$. Similarly, if $w \in \mathbb{R}^n$ and $1 \le k \le \ell \le n$, then $w_k^\ell = (w_k, \ldots, w_\ell) \in \mathbb{R}^{k-\ell+1}$. Occasionally we will write the weighted Hamming metric (1.3) as ρ_w to emphasize its dependence on w.

We use the indicator variable $\mathbf{1}_{\{\cdot\}}$ to assign 0-1 truth values to the predicate in $\{\cdot\}$. The ramp function is defined by $(z)_+ = z\mathbf{1}_{\{z>0\}}$. The positive reals are denoted by $\mathbb{R}_+ = (0, \infty)$.

The probability $\mathbf{P}$ and expectation $\mathbf{E}$ operators are defined with respect the measure space specified in context.

2 Linear programming inequality

The statement of our main requires a few preliminary definitions.

2.1 Definitions

Fix a finite set Ω, $n \in \mathbb{N}$ and $w \in \mathbb{R}_+^n$. We make the following definitions:

1. F_n denotes the set of all functions $g : \Omega^n \to \mathbb{R}$ (and $F_0 = \mathbb{R}$).
2. For $f \in F_n$, its *Lipschitz constant* with respect to ρ_w, denoted by $\|f\|_{\mathrm{Lip},w}$, is defined to be the smallest c for which
$$|f(x) - f(y)| \;\le\; c\rho_w(x,y), \qquad x, y \in \Omega^n;$$
any f with $\|f\|_{\mathrm{Lip},w} \le c$ is called c-Lipschitz.
3. For $v \in [0, \infty)$, define $\Phi_{w,n}^{+v} \subset F_n$ to be the set of all f such that $\|f\|_{\mathrm{Lip},w} \le 1$ and
$$0 \;\le\; f(x) \;\le\; \|w\|_1 + v, \qquad x \in \Omega^n;$$
we omit the $+v$ superscript when $v = 0$, writing simply $\Phi_{w,n}$. We use v as a "slack variable" to make the induction proof go through; in the applications, it is always 0.

4. The *projection operator* $(\cdot)'$ takes $g \in F_n$ to $g' \in F_{n-1}$ by

$$g'(z) = \sum_{x_1 \in \Omega} g(x_1 z), \qquad z \in \Omega^{n-1};$$

for $n = 1$, $g' \in F_0$ is the scalar $g' = \sum_{x_1 \in \Omega} g(x_1)$.

5. For $y \in \Omega$, the *y-section* operator $(\cdot)_y$ takes $g \in F_n$ to $g_y \in F_{n-1}$ by

$$g_y(x) = g(xy), \qquad x \in \Omega^{n-1};$$

for $n = 1$, $g_y \in F_0$ is the scalar $g(y)$.

6. The functional $\Psi_{w,n} : F_n \to \mathbb{R}$ is defined by $\Psi_{w,0}(\cdot) = 0$ and

$$\Psi_{w,n}(g) = w_1 \sum_{x \in \Omega^n} (g(x))_+ + \Psi_{w_2^n, n-1}(g'); \tag{2.1}$$

when $w_i \equiv 1$ we omit it from the subscript, writing simply Ψ_n. The letter Psi is a mnemonic for "Positive Summation, Iterated." Note that when g is non-negative, we have $\Psi_n(g) = \|w\|_1 \|g\|_1$.

7. The finite-dimensional vector space F_n is equipped with the inner product

$$\langle g, h \rangle = \sum_{x \in \Omega^n} g(x) h(x).$$

8. Two norms are defined on $g \in F_n$: the Φ_w-norm,

$$\|g\|_{\Phi,w} = \sup_{f \in \Phi_{w,n}} |\langle g, f \rangle| \tag{2.2}$$

and the Ψ_w-norm,

$$\|g\|_{\Psi,w} = \max_{s=\pm 1} \Psi_{w,n}(sg). \tag{2.3}$$

Remark 2.1. For the special case $w_i \equiv 1$, ρ_w is the unweighted Hamming metric used in [9]. It is straightforward to verify that Φ_w-norm and Ψ_w-norm satisfy the vector-space norm axioms for any $w \in \mathbb{R}_+^n$; this is done in [9] for $w_i \equiv 1$. Since we will not be appealing to any norm properties of these functionals, we omit the proof. Note that for any $y \in \Omega$, the projection and y-section operators commute; in other words, for $g \in F_{n+2}$, we have $(g')_y = (g_y)' \in F_n$ and so we may denote this common value by $g'_y \in F_n$:

$$g'_y(z) = \sum_{x_1 \in \Omega} g_y(x_1 z) = \sum_{x_1 \in \Omega} g(x_1 z y), \qquad z \in \Omega^n.$$

2.2 Statement and proof

The main result of this section is

Theorem 2.2. *For all $w \in \mathbb{R}^n_+$ and all $g \in F_n$, we have*

$$\|g\|_{\Phi,w} \quad \leq \quad \|g\|_{\Psi,w}\,.$$

Remark 2.3. This result is proved for $w_i \equiv 1$ in [9]. However, the proof given there is somewhat cumbersome and does not readily extend to the case of general w (in particular, it is not clear how to define the corresponding sub-Lipschitz polytopes).

The key technical lemma is a decomposition of $\Psi_{w,n}(\cdot)$ in terms of y-sections.

Lemma 2.4. *For all $n \geq 1$, $w \in \mathbb{R}^n_+$ and $g \in F_n$, we have*

$$\Psi_{w,n}(g) \quad = \quad \sum_{y\in\Omega}\left[\Psi_{w_1^{n-1},n-1}(g_y) + w_n\left(\sum_{x\in\Omega^{n-1}} g_y(x)\right)_+\right]. \tag{2.4}$$

Proof. We proceed by induction on n. To prove the $n = 1$ case, recall that Ω^0 is the set containing a single (null) word and that for $g \in F_1$, $g_y \in F_0$ is the scalar $g(y)$. Thus, by definition of $\Psi_{w,1}(\cdot)$, we have

$$\Psi_{w,1}(g) \quad = \quad w_1 \sum_{y\in\Omega} (g(y))_+\,,$$

which proves (2.4) for $n = 1$.

Suppose the claim holds for some $n = \ell \geq 1$. Pick any $w \in \mathbb{R}^{\ell+1}_+$ and $g \in F_{\ell+1}$ and examine

$$\begin{aligned}\sum_{y\in\Omega}\left[\Psi_{w_1^\ell,\ell}(g_y) + w_{\ell+1}\left(\sum_{x\in\Omega^\ell} g_y(x)\right)_+\right] &= \sum_{y\in\Omega}\left[\left(w_1\sum_{x\in\Omega^\ell}\Big(g_y(x)\Big)_+ + \Psi_{w_2^\ell,\ell-1}(g'_y)\right) + w_{\ell+1}\left(\sum_{x\in\Omega^\ell} g_y(x)\right)_+\right] \\ &= \sum_{y\in\Omega}\left[\Psi_{w_2^\ell,\ell-1}(g'_y) + w_{\ell+1}\left(\sum_{u\in\Omega^{\ell-1}} g'_y(u)\right)_+\right] + w_1\sum_{z\in\Omega^{\ell+1}}(g(z))_+\,,\end{aligned} \tag{2.5}$$

where the first equality follows from the definition of $\Psi_{w_1^\ell,\ell}$ in (2.1) and the second one from the straightforward identities

$$\sum_{y\in\Omega}\sum_{x\in\Omega^\ell}\Big(g_y(x)\Big)_+ \quad = \quad \sum_{z\in\Omega^{\ell+1}}(g(z))_+$$

and

$$\sum_{x\in\Omega^\ell} g_y(x) \quad = \quad \sum_{u\in\Omega^{\ell-1}} g'_y(u).$$

On the other hand, by definition we have

$$\Psi_{w,\ell+1}(g) \quad = \quad w_1\sum_{z\in\Omega^{\ell+1}}(g(z))_+ + \Psi_{w_2^{\ell+1},\ell}(g'). \tag{2.6}$$

To compare the r.h.s. of (2.5) with the r.h.s. of (2.6), note that the $w_1 \sum_{z\in\Omega^{\ell+1}} (g(z))_+$ term is common to both and

$$\sum_{y\in\Omega}\left[\Psi_{w_2^\ell,\ell-1}(g'_y) + w_{\ell+1}\left(\sum_{u\in\Omega^{\ell-1}} g'_y(u)\right)_+\right] \quad = \quad \Psi_{w_2^{\ell+1},\ell}(g')$$

by the inductive hypothesis. This establishes (2.4) for $n = \ell + 1$ and proves the claim. □

Our main result, Theorem 2.2, is an immediate consequence of

Theorem 2.5. *For all $n \geq 1$, $w \in \mathbb{R}_+^n$, $v \in [0,\infty)$ and $g \in F_n$, we have*

$$\sup_{f\in\Phi_{w,n}^{+v}} \langle g, f\rangle \quad \leq \quad \Psi_{w,n}(g) + v\left(\sum_{x\in\Omega^n} g(x)\right)_+ . \tag{2.7}$$

Proof. We will prove the claim by induction on n. For $n = 1$, pick any $w_1 \in \mathbb{R}_+$, $v \in [0,\infty)$ and $g \in F_1$. Since by construction any $f \in \Phi_{w_1,1}^{+v}$ is w_1-Lipschitz with respect to the discrete metric on Ω, f must be of the form

$$f(x) \quad = \quad \tilde{f}(x) + \tilde{v}, \qquad x \in \Omega,$$

where $\tilde{f} : \Omega \to [0, w_1]$ and $0 \leq \tilde{v} \leq v$ (in fact, we have the explicit value $\tilde{v} = (\max_{x\in\Omega} f(x) - w_1)_+$). Therefore,

$$\langle g, f\rangle \quad = \quad \langle g, \tilde{f}\rangle + \tilde{v}\sum_{x\in\Omega} g(x). \tag{2.8}$$

The first term in the r.h.s. of (2.8) is clearly maximized when $\tilde{f}(x) = w_1 \mathbf{1}_{\{g(x)>0\}}$ for all $x \in \Omega$, which shows that it is bounded by $\Psi_{w_1,1}(g)$. Since the second term in the r.h.s. of (2.8) is bounded by $v(\sum_{x\in\Omega} g(x))_+$, we have established (2.7) for $n = 1$.

Now suppose the claim holds for $n = \ell$, and pick any $w \in \mathbb{R}_+^{\ell+1}$, $v \in [0,\infty)$ and $g \in F_{\ell+1}$. By the reasoning given above (i.e., using the fact that $0 \leq f \leq v + \sum_{i=1}^{\ell+1} w_i$ and that f is 1-Lipschitz with respect to ρ_w), any $f \in \Phi_{w,\ell+1}^{+v}$, must be of the form $f = \tilde{f} + \tilde{v}$, where $\tilde{f} \in \Phi_{w,\ell+1}$ and $0 \leq \tilde{v} \leq v$. Thus we write $\langle g, f\rangle = \langle g, \tilde{f}\rangle + \tilde{v}\sum_{x\in\Omega^{\ell+1}} g(x)$ and decompose

$$\langle g, \tilde{f}\rangle \quad = \quad \sum_{y\in\Omega} \langle g_y, \tilde{f}_y\rangle, \tag{2.9}$$

making the obvious but crucial observation that

$$\tilde{f} \in \Phi_{w,\ell+1} \quad \Longrightarrow \quad \tilde{f}_y \in \Phi_{w_1^\ell,\ell}^{+w_{\ell+1}}.$$

Then it follows by the inductive hypothesis that

$$\langle g_y, \tilde{f}_y\rangle \quad \leq \quad \Psi_{w_1^\ell,\ell}(g_y) + w_{\ell+1}\left(\sum_{x\in\Omega^\ell} g_y(x)\right)_+ . \tag{2.10}$$

Applying Lemma 2.4 to (2.10), we have

$$\sum_{y\in\Omega}\langle g_y, \tilde{f}_y\rangle \;\le\; \sum_{y\in\Omega}\left[\Psi_{w_1^\ell,\ell}(g_y) + w_{\ell+1}\left(\sum_{x\in\Omega^\ell} g_y(x)\right)_+\right] \;=\; \Psi_{w,\ell+1}(g). \tag{2.11}$$

This, combined with (2.9) and the trivial bound

$$\tilde{v}\sum_{x\in\Omega^{\ell+1}} g(x) \;\le\; v\left(\sum_{x\in\Omega^{\ell+1}} g(x)\right)_+$$

proves the claim for $n = \ell + 1$ and hence for all n. □

Remark 2.6. We have given a method for bounding

$$\max_{\|f\|_{\mathrm{Lip},w}\le 1}\langle g, f\rangle$$

for a given $g \in \mathbb{R}^{\Omega^n}$. Note that the function being maximized, $F(\cdot) = \langle g, \cdot\rangle$ is linear in its argument and its domain is the finitely generated, compact, convex polytope $\Phi_{w,n} \subset \mathbb{R}^{\Omega^n}$ — hence the term "linear programming inequality." We make no use of this simple fact and therefore forgo its proof, but see [9, Lemma 4.4] for a proof of a closely related claim.

Remark 2.7. Although our technique bounds the value of a certain linear program, it apparently gives no hint as to the form of the solution. We hope to address this gap in future research. Another direction for future work is obtaining analogues of Theorem 2.2 for non-Hamming metrics on $\mathbb{R}^n$.

3 Applications

3.1 Azuma-Hoeffding martingale difference

This section assumes some familiarity with the notion of measure concentration; see References (in particular, [11, 12]) for introductory and survey material. Briefly, we shall concern ourselves with the metric probability space $(\Omega^n, \rho_w, \mathbf{P})$ where Ω is a finite set, $w \in \mathbb{R}_+^n$, ρ_w is the weighted Hamming metric defined in (1.3) and $\mathbf{P}$ is a (possibly non-product) probability measure on Ω^n. Our goal is to bound $\mathbf{P}\{|f - \mathbf{E}f| > t\}$ for suitable $f : \Omega^n \to \mathbb{R}$.

The method of martingale differences has been used to prove concentration of measure results since the work of Hoeffding, Azuma, and McDiarmid; see the exposition and references in [8, 9]. Let $(\Omega^n, \rho_w, \mathbf{P})$ be as defined above and associate to it the (canonical) random process $X = (X_i)_{1\le i\le n}$, $X_i \in \Omega$, satisfying

$$\mathbf{P}\{X \in A\} \;=\; \mathbf{P}(A)$$

for any $A \subset \Omega^n$.

For $1 \le i \le n$, $f : \Omega^n \to \mathbb{R}$ and $y_1^i \in \Omega^i$, define the *martingale difference*

$$V_i(f; y_1^i) \;=\; \mathbf{E}[f(X)\,|\,X_1^i = y_1^i] - \mathbf{E}[f(X)\,|\,X_1^{i-1} = y_1^{i-1}]. \tag{3.1}$$

Let

$$\bar{V}_i(f) \quad = \quad \max_{y_1^i \in \Omega^i} \left| V_i(f; y_1^i) \right| \tag{3.2}$$

and

$$D^2(f) \quad = \quad \sum_{i=1}^{n} \bar{V}_i^2(f).$$

Then Azuma's inequality [2] states that

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2\exp(-t^2/2D^2(f)) \tag{3.3}$$

(see [11] for a modern presentation and a short proof of (3.3)).

In [9] and [8], a technique was developed for bounding the martingale difference $V_i(f; y)$ in terms of the Lipschitz constant of f and mixing properties of the measure $\mathbf{P}$. To this end, the so-called η-mixing coefficients were introduced therein (see discussion ibid. regarding the appearance of these coefficients in earlier work of Marton [14] and Samson [19]).

For $1 \leq i < j \leq n$ and $x \in \Omega^i$, let

$$\mathcal{L}(X_j^n \,|\, X_1^i = x)$$

be the law (distribution) of X_j^n conditioned on $X_1^i = x$. For $y \in \Omega^{i-1}$ and $z, \dot{z} \in \Omega$, define

$$\eta_{ij}(y, z, \dot{z}) \quad = \quad \left\| \mathcal{L}(X_j^n \,|\, X_1^i = yz) - \mathcal{L}(X_j^n \,|\, X_1^i = y\dot{z}) \right\|_{\mathrm{TV}}, \tag{3.4}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm, defined here, for a signed measure τ on a finite space $\mathcal{X}$ by

$$\|\tau\|_{\mathrm{TV}} \quad = \quad \tfrac{1}{2} \sum_{x \in \mathcal{X}} |\tau(x)|.$$

If τ is *balanced* in the sense of $\sum_{x \in \mathcal{X}} \tau(x) = 0$, we further have

$$\|\tau\|_{\mathrm{TV}} = \sum_{x \in \mathcal{X}} (\tau(x))_+ . \tag{3.5}$$

Additionally, define

$$\bar{\eta}_{ij} \quad = \quad \max_{y \in \Omega^{i-1}} \max_{z, \dot{z} \in \Omega} \eta_{ij}(y, z, \dot{z}).$$

The main application of Theorem 2.5 to measure concentration is the following bound on the martingale difference:

Theorem 3.1. *Let Ω be a finite set, and let $(X_i)_{1 \leq i \leq n}$, $X_i \in \Omega$ be the random process associated with the measure $\mathbf{P}$ on Ω^n. Let Δ_n be the upper-triangular $n \times n$ matrix defined by $(\Delta_n)_{ii} = 1$ and*

$$(\Delta_n)_{ij} = \bar{\eta}_{ij}, \qquad 1 \leq i < j \leq n. \tag{3.6}$$

Then, for all $w \in \mathbb{R}^n_+$ and $f : \Omega^n \to \mathbb{R}$, we have

$$\sum_{i=1}^{n} \bar{V}_i^2(f) \quad \leq \quad \|f\|^2_{\mathrm{Lip},w} \|\Delta_n w\|_2^2, \tag{3.7}$$

where $\bar{V}_i^2(f)$ is defined in (3.2).

Remark: Since $\bar{V}_i(f)$ and $\|f\|_{\mathrm{Lip},w}$ are both homogeneous functionals of f (in the sense that $T(af) = |a|T(f)$ for $a \in \mathbb{R}$), there is no loss of generality in taking $\|f\|_{\mathrm{Lip},w} = 1$. Additionally, since $V_i(f; y)$ is translation-invariant (in the sense that $V_i(f; y) = V_i(f + a; y)$ for all $a \in \mathbb{R}$), there is no loss of generality in restricting the range of f to $[0, \mathrm{diam}_{\rho_w}(\Omega^n)]$. In other words, it suffices to consider $f \in \Phi_{w,n}$. The proof will follow closely that of [9, Theorem 5.1]. The extension of this result to countable Ω is quite straightforward, along the lines of [9, Lemma 6.1].

Proof. We must show that for any $1 \leq i < n$ and $f \in \Phi_{w,n}$,

$$\bar{V}_i(f) \leq w_i + \sum_{j=i+1}^{n} w_j \bar{\eta}_{ij} = \sum_{j=1}^{n} (\Delta_n)_{ij} w_j = (\Delta_n w)_i, \tag{3.8}$$

whence (3.7) follows immediately by squaring and summing over i. We begin by invoking [9, Lemma 3.1], which bounds $\bar{V}_i(f)$ by a related functional:

$$\bar{V}_i(f) \leq \max_{y^{i-1} \in \Omega^{i-1}, z, \dot{z} \in \Omega} \left| \hat{V}_i(f; y^{i-1}, z, \dot{z}) \right|,$$

where, for $y^{i-1} \in \Omega^{i-1}$ and $z, \dot{z} \in \Omega$,

$$\hat{V}_i(f; y^{i-1}, z, \dot{z}) = \mathbf{E}[f(X) \,|\, X^i = y^{i-1} z] - \mathbf{E}[f(X) \,|\, X^i = y^{i-1} \dot{z}].$$

Next, we observe that the functional $f \mapsto \hat{V}_i(f; y^{i-1}, z, \dot{z})$ is given by

$$\hat{V}_i(f; y^{i-1}, z, \dot{z}) = \left\langle g_{y^{i-1}, z, \dot{z}}, f \right\rangle,$$

where $g_{y^{i-1}, z, \dot{z}} \in F_n$ is defined by

$$g_{y^{i-1}, z, \dot{z}}(x) = \mathbf{1}_{\{x^{i-1} = y^{i-1}\}} \left(\mathbf{1}_{\{x_i = z\}} \mathbf{P}(x_{i+1}^n \,|\, y^{i-1} z) - \mathbf{1}_{\{x_i = \dot{z}\}} \mathbf{P}(x_{i+1}^n \,|\, y^{i-1} \dot{z}) \right)$$

and $\mathbf{P}(x_{i+1}^n \,|\, y^{i-1} z)$ is a shorthand for $\mathbf{P}(X_{i+1}^n = x_{i+1}^n \,|\, X_1^{i-1} = y^{i-1} z)$. We further notice that the structure of $g_{y^{i-1}, z, \dot{z}}$ implies that

$$\left\langle g_{y^{i-1}, z, \dot{z}}, f \right\rangle = \left\langle T_{y^{i-1}}[g_{y^{i-1}, z, \dot{z}}], T_{y^{i-1}}[f] \right\rangle,$$

where, for $t \in \Omega^{i-1}$, the operator $T_t : F_n \to F_{n-i+1}$ is defined by

$$T_t[h](x) = h(tx), \qquad x \in \Omega^{n-i+1}.$$

Applying Theorem 2.5, we get

$$\left\langle T_{y^{i-1}}[g_{y^{i-1}, z, \dot{z}}], T_{y^{i-1}}[f] \right\rangle \quad \leq \quad \Psi_{w_i^n, n-i+1}(T_{y^{i-1}}[g_{y^{i-1}, z, \dot{z}}]).$$

Thus, in order to prove (3.8), it suffices to show that

$$\Psi_{w_i^n,n-i+1}(g^{(L)}) \le w_i + \sum_{j=i+1}^{n} w_j \bar{\eta}_{ij},$$

where $L = n-i+1$ and $g^{(L)} = T_{y^{i-1}}[g_{y^{i-1},z,\dot{z}}]$. For $\ell = L, L-1, \ldots, 2$, define

$$g^{(\ell-1)} = (g^{(\ell)})',$$

where $(')$ is the marginal projection operator defined in Section 2.1. Then $g^{(\ell)} \in F_\ell$ and a direct calculation shows that for $i < j \le n$,

$$g^{(n-j+1)}(x) \;=\; \mathbf{P}\Big\{X_j^n = x | X^i = y^{i-1}z\Big\} - \mathbf{P}\Big\{X_j^n = x | X^i = y^{i-1}\dot{z}\Big\}, \qquad x \in \Omega^{n-j+1}. \tag{3.9}$$

Since $g^{(n-j+1)}$ is a difference of two probability measures on Ω^{n-j+1}, we have by (3.5) that

$$\left\|g^{(n-j+1)}\right\|_{\mathrm{TV}} = \sum_{x \in \Omega^{n-j+1}} \left(g^{(n-j+1)}(x)\right)_+ .$$

Together with (3.9), this immediately shows that for $i < j \le n$,

$$\sum_{x \in \Omega^{n-j+1}} \left(g^{(n-j+1)}(x)\right)_+ = \eta_{ij}(y^{i-1}, z, \dot{z}). \tag{3.10}$$

Now, from the definition of the $\Psi_{w,n}$ functional (2.1), we see that

$$\Psi_{w_i^n,n-i+1}(g^{(L)}) \quad = \quad w_i \sum_{x \in \Omega^L} \left(g^{(L)}(x)\right)_+ + \Psi_{w_{i+1}^n,n-i}(g^{(L-1)}).$$

Continuing to unravel the recursion in (2.1) and combining with (3.10), we obtain (3.8).

□

Corollary 3.2. *Let Ω be a finite set and $\mathbf{P}$ a measure on Ω^n, for $n \ge 1$. For any $w \in \mathbb{R}_+^n$ and $f : \Omega^n \to \mathbb{R}$, we have*

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \le \quad 2\exp\left(-\frac{t^2}{2\|f\|_{\mathrm{Lip},w}^2 \|\Delta_n w\|_2^2}\right).$$

To place our results in context, let us recall some classic bounds. McDiarmid's inequality [15] may be stated as

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \le \quad 2\exp\left(-\frac{2t^2}{\|f\|_{\mathrm{Lip},w}^2 \|w\|_2^2}\right), \tag{3.11}$$

where $\mathbf{P}$ is a product measure. Marton's result [13] states that if $f : \Omega^n \to \mathbb{R}$ is a 1-Lipschitz function with respect to ρ_w, $w_i \equiv n^{-1}$, and $\mathbf{P}$ is a contracting homogeneous Markov chain with Doeblin coefficient $\theta < 1$ then

$$\mathbf{P}\Big\{\big|f - M_f\big| > t\Big\} \quad \le \quad 2\exp\left[-2n\left(t(1-\theta) - \sqrt{\frac{\log 2}{2n}}\right)^2\right], \tag{3.12}$$

where M_f is a $\mathbf{P}$-median of f.

Remark 3.3. It may be shown via Markov's contraction lemma [8, Lemma 4.1.1] that for contracting Markov chains with Doeblin coefficient θ, we have

$$(\Delta_n)_{ij} \leq \theta^{j-i}$$

for $1 \leq i < j \leq n$. Following [5], we use Young's inequality to obtain

$$\|\Delta_n w\|_2^2 \leq (1-\theta)^{-2} \|w\|_2^2$$

for contracting Markov chains. Thus, we have that if $f : \Omega^n \to \mathbb{R}$ is a 1-Lipschitz function with respect to any ρ_w, and $\mathbf{P}$ is a contracting (possibly inhomogeneous) Markov chain with Doeblin coefficient bounded by $\theta < 1$, then

$$\mathbf{P}\{|f - \mathbf{E}f| > t\} \quad \leq \quad 2\exp\left(-\frac{(1-\theta)^2 t^2}{2\|w\|_2^2}\right). \tag{3.13}$$

The contraction method is extended to more general processes in [10].

Note that (3.13) generalizes (3.12) (up to constants and decaying terms in the exponent) both over the metrics ($w \in \mathbb{R}_+^n$ is arbitrary) and the class of measures (homogeneity and strict contractivity are not required); see [8, 9] for details. Analogous results have recently been obtained by [4] and [5]. Note further that all of our concentration results continue to hold if the metric ρ_w defined in (1.3) is replaced by an ℓ_1 sum of arbitrary discrete metrics on Ω:

$$\tilde{\rho}(x,y) = \sum_{i=1}^n d_i(x_i, y_i), \qquad x, y \in \Omega^n.$$

Taking $w_i = \operatorname{diam}(\Omega, d_i)$, it is easy to see that any function $F : \Omega^n \to \mathbb{R}$ that is L-Lipschitz with respect to $\tilde{\rho}$ is also L-Lipschitz with respect to ρ_w. Thus, Corollary 3.2 sharpens Theorem 1 of [17], where the stronger γ-mixing assumption is made.

3.2 Marton's transportation inequality

In [13], Marton developed the powerful *transportation* method for proving concentration inequalities. Let Ω^n be equipped with the metric ρ defined in (1.3). For two distributions μ and ν on Ω^n, define also the *relative entropy* (or Kullback-Leibler divergence) of ν with respect to μ as

$$H(\nu|\mu) \;=\; \mathrm{Ent}_\mu\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \;=\; \int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \mathrm{d}\nu$$

whenever $\nu \ll \mu$ with Radon-Nikodým derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

The measure μ is said to satisfy a transportation inequality with constant $a > 0$ if

$$T_\rho(\mu, \nu) \quad \leq \quad \frac{1}{a}\left[\frac{1}{2n} H(\nu|\mu)\right]^{1/2} \tag{3.14}$$

for every ν, where T_ρ is defined in (1.1).

This condition implies concentration for μ:

$$\mu\left\{x \in \Omega^n : \left|f - M_f\right| > t\right\} \quad \leq \quad 2\exp\left[-2n\left(at - \sqrt{\frac{\log 2}{2n}}\right)^2\right] \tag{3.15}$$

provided $t \geq \sqrt{\log 2/(2n)}$, where M_f is a μ-median of f (see [11] or [13] for a simple derivation of this result).

A consequence of Kantorovich duality (1.2) is that to bound the transportation cost $T_\rho(\mu,\nu)$ it suffices to bound $\sup_{\|f\|_{\text{Lip}}\leq 1}\langle \mu-\nu, f\rangle$. But this is precisely what Theorem 2.2 accomplishes:

Corollary 3.4. *If Ω is a finite set equipped with the weighted Hamming metric ρ as in (1.3) and μ,ν are distributions on Ω^n, then*

$$T_\rho(\mu,\nu) \quad \leq \quad \Psi_{w,n}(\mu-\nu).$$

The implications for concentration of measure are that in cases where bounding T_ρ directly (e.g., via coupling methods as in [13]) is difficult, one could attempt to prove an inequality of the type

$$\Psi_{w,n}(\mu-\nu) \quad \leq \quad \frac{1}{a}\left[\frac{1}{2n}H(\nu|\mu)\right]^{1/2}, \tag{3.16}$$

which a fortiori would imply (3.14). This approach has the advantage that $\Psi_{w,n}(\mu-\nu)$ is readily computable by an explicit, closed-form formula, while bounding T_ρ typically involves discovering clever couplings. A drawback of this approach is that the resulting bounds are likely to be cruder than direct methods.

Acknowledgments

The author thanks the referee for carefully reading the manuscript and providing insightful comments. Partial support was provided by the ISRAEL SCIENCE FOUNDATION (grant No. 1141/12).

References

[1] Aaron Archer, Jittat Fakcharoenphol, Chris Harrelson, Robert Krauthgamer, Kunal Talwar, and Éva Tardos. Approximate classification via earthmover metrics. In *SODA*, pages 1079–1087, 2004.

[2] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. Journal*, 19:357–367, 1967.

[3] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, New York, NY, USA, 2002. ACM.

[4] Jean-René Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1-2):201–225, 2007.

[5] Jean-René Chazottes and Frank Redig. Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.*, 14:no. 40, 1162–1180, 2009.

[6] Chandra Chekuri, Sanjeev Khanna, Joseph Naor, and Leonid Zosin. Approximation algorithms for the metric labeling problem via a new linear programming formulation. In *SODA*, pages 109–118, 2001.

[7] Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *ICCV*, 2003.

[8] Leonid (Aryeh) Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

[9] Leonid (Aryeh) Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.*, 36(6):2126–2158, 2008.

[10] Aryeh Kontorovich. Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields*, 2012+.

[11] Michel Ledoux. *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs Vol. 89*. American Mathematical Society, 2001.

[12] Gábor Lugosi. Concentration-of-measure inequalities, `http://www.econ.upf.es/~lugosi/anu.ps`. 2003.

[13] Katalin Marton. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996.

[14] Katalin Marton. Measure concentration and strong mixing. *Studia Scientiarum Mathematicarum Hungarica*, 19(1-2):95–113, 2003.

[15] Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, volume 141 of LMS Lecture Notes Series*, pages 148–188. Morgan Kaufmann Publishers, San Mateo, CA, 1989.

[16] Shmuel Peleg, Michael Werman, and Hillel Rom. A unified approach to the change of resolution: Space and gray-level. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):739–742, 1989.

[17] Emmanuel Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908, 2000.

[18] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[19] Paul-Marie Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[20] L. N. Vasershtein. Markov processes over denumerable products of spaces describing large system of automata. *Problemy Peredači Informacii*, 5(3):64–72, 1969.

[21] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[22] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009.