Information Constrained Optimal Transport: From Talagrand, to Marton, to Cover

Yikun Bai
University of Delaware
Email: bai@udel.edu

Xiugang Wu
University of Delaware
Email: xwu@udel.edu

Ayfer Özgür
Stanford University
Email: aozgur@stanford.edu

***Abstract*—The optimal transport problem studies how to transport one measure to another in the most cost-effective way and has wide range of applications from economics to machine learning. In this paper, we introduce and study an information constrained variation of this problem. Our study yields a strengthening and generalization of Talagrand's celebrated transportation cost inequality. Following Marton's approach, we show that the new transportation cost inequality can be used to recover old and new concentration of measure results. Finally, we provide an application of this inequality to network information theory. We show that it can be used to recover a recent solution to a long-standing open problem posed by Cover regarding the capacity of the relay channel.**

I. Introduction

The optimal transport (OT) theory, pioneered by Monge [1] and Kantorovich [2], studies how to distribute supply to meet demand in the most cost-effective way. It has many known connections with, and applications to areas such as geometry, quantum mechanics, fluid dynamics, optics, mathematical statistics, and meteorology. More recently, it has received renewed interest due to its increasingly many applications in imaging sciences, computer vision and machine learning.

A. Optimal Transport Problem

The basic OT problem in Kantorovich's probabilistic formulation can be described as follows. Let $\mathcal{Z}$ and $\mathcal{Y}$ be two measurable spaces, $\mathcal{P}(\mathcal{Z})$ and $\mathcal{P}(\mathcal{Y})$ be the sets of all probability measures on $\mathcal{Z}$ and $\mathcal{Y}$ respectively, and $\mathcal{P}(\mathcal{Z}\times\mathcal{Y})$ be the set of all joint probability measures on $\mathcal{Z}\times\mathcal{Y}$. Let $c:\mathcal{Z}\times\mathcal{Y}\to\mathbb{R}_+$ be a non-negative measurable function, which is called the cost function. Given two probability measures $P_Z\in\mathcal{P}(\mathcal{Z})$ and $P_Y\in\mathcal{P}(\mathcal{Y})$, the set of couplings of P_Z and P_Y, denoted by $\Pi(P_Z,P_Y)$, refers to the set of all joint probability measures $P\in\mathcal{P}(\mathcal{Z}\times\mathcal{Y})$ such that their marginal measures are P_Z and P_Y. The OT problem is to find the coupling P in $\Pi(P_Z,P_Y)$ that minimizes the expected cost:

$$\inf_{P\in\Pi(P_Z,P_Y)} \mathrm{E}_P[c(Z,Y)]. \tag{1}$$

A special case of particular interest is when $\mathcal{Z}=\mathcal{Y}=\mathbb{R}$ and $c(z,y)=|z-y|^p$, in which case the quantity

$$W_p(P_Z,P_Y)\triangleq \inf_{P\in\Pi(P_Z,P_Y)} \{\mathrm{E}_P[|Z-Y|^p]\}^{1/p} \tag{2}$$

The work of A. Ozgur was partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center under grant agreement CCF-0939370, and NSF award CCF-1704624.

defines a distance between two probability measures P_Z and P_Y and is called the p-th order Wasserstein distance. Various transportation cost inequalities have been developed that upper bound the Wasserstein distance between two measures P_Z and P_Y. For example, the celebrated Talagrand's transportation inequality [3] states that

$$W_2^2(P_Z,P_Y)\le 2D(P_Z\|P_Y) \tag{3}$$

when P_Y is standard Gaussian $\mathcal{N}(0,1)$ and $P_Z\ll P_Y$.

B. Information Constrained Optimal Transport

In this paper, we introduce and study a variation of the OT problem which we call the information constrained OT problem. Here, we want to find the coupling P in $\Pi(P_Z,P_Y)$ that minimizes the expected cost while ensuring that the mutual information $I_P(Z;Y)$ between Z and Y under the coupling P does not exceed some pre-specified value R:

$$\inf_{P\in\Pi(P_Z,P_Y):I_P(Z;Y)\le R} \mathrm{E}_P[c(Z,Y)]. \tag{4}$$

It is worth mentioning that an equivalent formulation to problem (4) has received significant recent interest in the machine learning literature, where one seeks to minimize the cost-information Lagrangian:

$$\inf_{P\in\Pi(P_Z,P_Y)} \{\mathrm{E}_P[c(Z,Y)]+\lambda I_P(Z;Y)\}. \tag{5}$$

The problem (5) generally appears under the name entropy regularized OT or Sinkhorn distances in the machine learning literature. This interest in (5) has been mainly motivated by computational considerations; in many cases computing the regularized OT in (5) from data turns out to be easier than computing the OT in (1), which motivates the use of (5) instead of (1) as a distance [4]. For certain inference tasks, (5) also appears to be a more suitable distance than (1) leading to superior empirical performance [5]. In contrast, in this paper we are interested in understanding the solution of the problem (4) as well as its fundamental connections to concentration of measure and network information theory.

C. Summary of Results

In the information constrained OT setup, one can similarly define the Wasserstein distance between two measures P_Z and P_Y subject to the information constraint R:

$$W_p(P_Z,P_Y;R)\triangleq \inf_{\substack{P\in\Pi(P_Z,P_Y):\\ I_P(Z;Y)\le R}} \{\mathrm{E}_P[|Z-Y|^p]\}^{1/p}. \tag{6}$$

Note that when $R = \infty$, (6) reduces to the unconstrained Wasserstein distance in (2). The main result of this paper, proved in Section II, is an upper bound on $W_2(P_Z, P_Y; R)$ for any $R \in \mathbb{R}_+$ when P_Y is standard Gaussian and $P_Z \ll P_Y$:

$$W_2^2(P_Z, P_Y; R) \le \mathrm{E}[Z^2] + 1 - 2\sqrt{\frac{1}{2\pi e} e^{2h(Z)} \left(1 - e^{-2R}\right)}. \tag{7}$$

This new transportation inequality sharpens and generalizes Talagrand's inequality in (3). Indeed, by setting R to be ∞ in (7), one can obtain a sharpened bound on the unconstrained Wasserstein distance:

$$W_2^2(P_Z, P_Y) \le \mathrm{E}[Z^2] + 1 - 2\sqrt{\frac{1}{2\pi e} e^{2h(Z)}}. \tag{8}$$

It is easy to check that the R.H.S. of (8) is smaller than or equal to that of Talagrand's inequality in (3) for any P_Z, and therefore (8) is uniformly tighter than (3). Moreover, the new inequality (7) captures the trade-off between information and transportation cost, which goes beyond the scope of (3). This trade-off turns out to be tight when P_Z is Gaussian.

Since the pioneering work of Marton [6], [7], it has been known that Talagrand's transportation inequality (3) captures essentially the same geometric phenomenon as the Gaussian isoperimetric inequality, both of which can be used to derive concentration of measure in Gaussian space. What are the geometric implications of the new transportation inequalities in (7) and (8)? In Section III, we show that the strengthening (8) of Talagrand's inequality can be used to prove concentration of measure on the sphere, which can be shown to imply concentration of measure in Gaussian space. In other words, (8) captures a stronger isoperimetric phenomenon than (3), the one on the sphere rather than that in Gaussian space. Furthermore, we show in Section III that the information constrained transportation inequality in (7) captures a new isoperimetric phenomenon on the sphere that has not been known before the recent work [8], co-authored by a subset of the authors. Different from the standard isoperimetric inequality on the sphere where one is interested in the extremal set that minimizes the measure of its neighborhood, this new isoperimetric result deals with the set that has minimal intersection measure with the neighborhood of a randomly chosen point on the sphere.

Finally, in Section IV we demonstrate an application of the new transportation inequality (7) to network information theory. We use it to recover the solution of a problem posed by Cover, "The Capacity of the Relay Channel", in *Open Problems in Communication and Computation*, Springer-Verlag, 1987, in the canonical Gaussian case. This problem was recently solved in the Gaussian case in [8], [9]. The proof in [8], [9] relied on intricate geometric arguments based on typical sets, while (7) allows us to recover the same result almost immediately.

II. New Transportation Inequalities

Before stating and proving our new transportation inequalities, let us first formalize the definition of the Wasserstein distance and Talagrand's transportation inequality; see also [10]. Let (Ω, d) be a Polish metric space. Given $p \ge 1$, let $\mathcal{P}_p(\Omega)$ denote the space of all Borel probability measures ν on Ω such that the moment bound $\mathrm{E}_{\omega\sim\nu}[d^p(\omega, \omega_0)] < \infty$ holds for some (and hence all) $\omega_0 \in \Omega$.

Definition 2.1 (Wasserstein Distance): The p-th order Wasserstein distance between $P_Z, P_Y \in \mathcal{P}_p(\Omega)$ is defined as

$$W_p(P_Z, P_Y) \triangleq \inf_{P \in \Pi(P_Z, P_Y)} \left\{\mathrm{E}_P[d^p(Z, Y)]\right\}^{1/p}. \tag{9}$$

If $p = 2$, $\Omega = \mathbb{R}$ with $d(z, y) = |z - y|$, and P_Y is atomless, then the optimal coupling that achieves the infimum in (9) is given by the deterministic mapping $Z = F_Z^{-1} \circ F_Y(Y)$ where F_Y is the cdf of P_Y, i.e. $F_Y(y) = P_Y(Y \le y)$ and F_Z^{-1} is the quantile function of P_Z, i.e. $F_Z^{-1}(\alpha) = \inf\{z \in \mathbb{R} : F_Z(z) \ge \alpha\}$. For convenience, in this paper we denote the mapping $F_Z^{-1} \circ F_Y$ by g, and call it the increasing rearrangement function. Building on this optimal coupling and tensorization [10], one can prove the following result for the case when $\Omega = \mathbb{R}^n$ and $d(z^n, y^n) = \|z^n - y^n\|_2$, known as Talagrand's transportation inequality.

Proposition 2.1 (Talagrand, 96): For two probability measures $P_{Z^n} \ll P_{Y^n}$ on $\mathbb{R}^n$ with $P_{Y^n} = \mathcal{N}(0, I_n)$,

$$W_2^2(P_{Z^n}, P_{Y^n}) \le 2D(P_{Z^n} \| P_{Y^n}), \tag{10}$$

where the inequality is tight if and only if P_{Z^n} is a shifted version of P_{Y^n}, i.e. $P_{Z^n} = \mathcal{N}(\mu, I_n)$ for some $\mu \in \mathbb{R}^n$.

A. Sharpening Talagrand's Transportation Inequality

Talagrand's transportation inequality can be sharpened to the following; see also [11], [12] for related results.

Theorem 2.1: For $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} \ll P_{Y^n}$,

$$W_2^2(P_{Z^n}, P_{Y^n}) \le \mathrm{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n}h(Z^n)}}, \tag{11}$$

where the inequality is tight when P_{Z^n} is isotropic Gaussian, i.e. $P_{Z^n} = \mathcal{N}(\mu, \sigma^2 I_n)$ for some $\mu \in \mathbb{R}^n$ and $\sigma > 0$.

Note that compared to Talagrand's transportation inequality, which is tight only when $P_{Z^n} = \mathcal{N}(\mu, I_n)$, the upper bound of the Wasserstein distance in Thm. 2.1 is tight for a wider class of P_{Z^n}, i.e. when P_{Z^n} is isotropic Gaussian. If fact, it can been shown that this transportation inequality is in general stronger than Talagrand's, i.e. R.H.S. of (11) $\le$ R.H.S. of (10), for any $P_{Z^n} \ll P_{Y^n}$ where the inequality holds with equality iff $h(Z^n) = \frac{n}{2} \ln 2\pi e$.

B. Extension to Information Constrained OT

The transportation inequality in Thm. 2.1 can be extended to the information constrained case.

Definition 2.2 (Information Constrained Wasserstein Distance): The p-th order Wasserstein distance between $P_Z, P_Y \in \mathcal{P}_p(\Omega)$ subject to information constraint R is defined as

$$W_p(P_Z, P_Y; R) \triangleq \inf_{P \in \Pi(P_Z, P_Y): I_P(Z;Y) \le R} \left\{\mathrm{E}_P[d^p(Z, Y)]\right\}^{1/p}.$$

For the case when $\Omega = \mathbb{R}^n$ and $d(z^n, y^n) = \|z^n - y^n\|_2$, we can prove the following bound on the information constrained Wasserstein distance.

Theorem 2.2: For $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} \ll P_{Y^n}$,

$$W_2^2(P_{Z^n}, P_{Y^n}; R)$$
$$\leq \mathrm{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n}h(Z^n)} \left(1 - e^{-\frac{2R}{n}}\right)}. \quad (12)$$

The above theorem characterizes a trade-off between the Wasserstein distance and the information constraint, as depicted in Fig. 1. This includes Thm. 2.1 as an extreme case by letting $R \to \infty$. The other extreme case is when $R = 0$, where now Z^n and Y^n are forced to be independent, and therefore the information constrained Wasserstein distance simply reduces to $\mathrm{E}[\|Z^n\|^2] + n$.

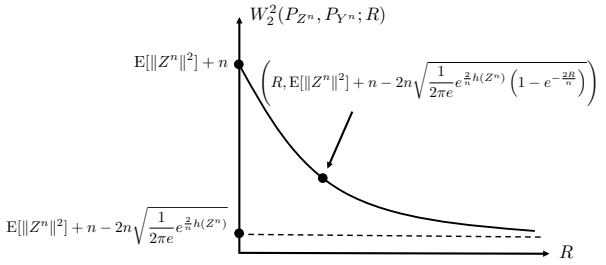

Fig. 1. Wasserstein distance-information constraint tradeoff.

Inequality (12) can be shown to be tight when P_{Z^n} is isotropic Gaussian; i.e., when $P_{Z^n} = \mathcal{N}(\mu, \sigma^2 I_n)$ for some μ and σ^2, the inequality in (12) is achieved with equality. Therefore, the trade-off characterized in Thm. 2.2 is indeed fundamental when P_{Z^n} is isotropic Gaussian.

C. Proof of Our Transportation Inequalities

We now prove the transportation inequalities stated in Thms. 2.1–2.2. Since Thm. 2.2 includes Thm. 2.1 as special case, it suffices to prove Thm. 2.2. Due to page limit, here we focus on proving the $n = 1$ case, i.e. inequality (7); the general n-dimensional case (12) can be obtained via tensorization [10].

To show (7), it suffices to construct a coupling P of P_Z and P_Y such that the information constraint $I_P(Z;Y) \leq R$ is satisfied and simultaneously $\mathrm{E}_P[(Z-Y)^2]$ is bounded by the R.H.S. of (7). For this, let $Y = \sqrt{1 - e^{-2R}} Y_1 + e^{-R} Y_2$, where $Y_1, Y_2 \sim \mathcal{N}(0,1)$ are two independent standard Gaussian random variables, and let Z be defined by $Z = g(Y_1) = F_Z^{-1} \circ F_{Y_1}(Y_1)$. It is easy to verify that the joint distribution P of (Z,Y) defined by the above is indeed a coupling of P_Z and P_Y. To see that the this coupling satisfies the information constraint, note that

$$\begin{aligned} I_P(Z;Y) &= h(Y) - h(\sqrt{1 - e^{-2R}} Y_1 + e^{-R} Y_2 | Z) \\ &= h(Y) - h(e^{-R} Y_2 | Z) \quad (13) \\ &= h(Y) - h(e^{-R} Y_2) \quad (14) \\ &= R \end{aligned}$$

where (13) holds because g is a one-to-one mapping and thus Y_1 is determined given Z, and (14) follows from the independence between Y_2 and Z.

Moreover, with the construction $Z = g(Y_1)$ we have

$$\begin{aligned} \mathrm{E}[Y_1 Z] &= \mathrm{E}[Y_1 g(Y_1)] \\ &= \int y_1 g(y_1) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_1^2}{2}\right) dy_1 \\ &= \int g'(y_1) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_1^2}{2}\right) dy_1 \quad (15) \\ &= \mathrm{E}[g'(Y_1)] \\ &= \mathrm{E}\left[\frac{f_{Y_1}(Y_1)}{f_Z(g(Y_1))}\right] \quad (16) \\ &\geq \exp\left(\mathrm{E}\left[\ln \frac{f_{Y_1}(Y_1)}{f_Z(g(Y_1))}\right]\right) \quad (17) \\ &= \exp(h(Z) - h(Y_1)) \\ &= \sqrt{\frac{1}{2\pi e} e^{2h(Z)}} \end{aligned}$$

where (15) follows from integration by part, (16) holds because $f_{Y_1}(y_1) = \frac{d}{dy_1} F_Z(g(y_1)) = f_Z(g(y_1)) g'(y_1)$ and (17) follows from Jensen's inequality. Therefore, $\mathrm{E}_P[(Z-Y)^2]$ can be upper bounded by

$$\begin{aligned} \mathrm{E}_P[(Z-Y)^2] &= \mathrm{E}[Z^2] + 1 - 2\mathrm{E}_P[YZ] \\ &= \mathrm{E}[Z^2] + 1 - 2\sqrt{1 - e^{-2R}} \mathrm{E}[Y_1 Z] \\ &\leq \mathrm{E}[Z^2] + 1 - 2\sqrt{1 - e^{-2R}} \sqrt{\frac{1}{2\pi e} e^{2h(Z)}} \\ &= \text{R.H.S. of (7).} \end{aligned}$$

This completes the proof of Thm. 2.2 in the $n = 1$ case.

III. Geometry: Concentration and Isoperimetry

Transportation cost inequalities of the form (3) are known to imply concentration of measure, an inherently geometric phenomenon tightly coupled with isoperimetric inequalites. This section discusses the geometric implications of Theorems 2.1–2.2. [1] For this, we begin with the geometry of Talagrand's transportation inequality.

A. Concentration and Isoperimetry in Gaussian Space

Consider a high-dimensional[2] Euclidean space $\mathbb{R}^m$. For any $A \in \mathbb{R}^m$ and $t > 0$, let A_t denote the t-blowup set of A:

$$A_t = \{x^m \in \mathbb{R}^m : \|x^m - a^m\| \leq t \text{ for some } a^m \in A\}.$$

The following concentration of measure result is generally known as the blowing-up lemma in Gaussian space [10].

Proposition 3.1: Let P_{Y^m} be the standard Gaussian measure on $\mathbb{R}^m$. For any $A \in \mathbb{R}^m$ with $P_{Y^m}(A) \geq e^{-ma}$,

$$P_{Y^m}(A_t) \to 1 \text{ as } m \to \infty$$

when $t \geq \sqrt{2m(a + \epsilon)}$ for some $\epsilon > 0$.

Roughly, the above result states that under the product Gaussian measure, slightly blowing up any set with a small but exponentially significant probability suffices to increase its probability to nearly 1; hence the name blowing-up lemma.

[1] Due to space constraints, we limit the presentation to formal statements of the results and informal discussions of their connections, delegating proofs to the long version of the paper.

[2] Here we use m instead of n to denote the dimension, since in this section m scales to infinity while in the previous sections the dimension n is fixed.

This lemma can be thought of as a consequence of the isoperimetric inequality in Gaussian space, which says that among all sets with the same Gaussian measure, a halfspace minimizes the measure of its t-blowup. Therefore, if we start with two sets A and H, where $P_{Y^m}(A) = P_{Y^m}(H)$ and H is a halfspace, then $P_{Y^m}(A_t) \geq P_{Y^m}(H_t)$ and hence it suffices to check that $P_{Y^m}(H_t) \to 1$, which follows from a simple calculation.

An alternative approach to proving the above blowing-up lemma, pioneered by Marton [6], [7], is through Talagrand's transportation inequality. A formal proof via this approach can be found in [10]. To get a feel for the connection between these two seemingly disjoint results, recall that Talagrand's inequality in (3) asserts that there exists a joint distribution P of (Z, Y) such that $\mathrm{E}_P[(Z-Y)^2] \leq 2D(P_Z \| P_Y)$, for $P_Y = \mathcal{N}(0,1)$ and $P_Z \ll P_Y$. Therefore, if we generate (Z^m, Y^m) i.i.d. according to P, then by the law of large numbers Z^m and Y^m are within distance $\sqrt{2mD(P_Z\|P_Y)}$ with high probability (w.h.p.), i.e.

$$\frac{1}{m}\|Z^m - Y^m\|_2^2 \to \mathrm{E}_P[(Z-Y)^2] \leq 2D(P_Z\|P_Y). \quad (18)$$

Roughly speaking, this allows us to control the distance between the typical set of Z^m, call it A, and Y^m, and therefore how much A needs to be blown-up to have probability approaching 1 under the measure of Y^m.

B. Concentration and Isoperimetry on the Sphere

We now show that transportation inequality (8) also has interesting geometric consequences. In particular, it implies the following concentration result on the sphere: Let Y^m be uniformly distributed on the unit sphere $\mathbb{S}^{m-1} \subseteq \mathbb{R}^m$. A spherical cap with angle θ is defined as a ball on $\mathbb{S}^{m-1}$ in the angle $\angle(z^m, y^m) = \arccos(\langle z^m, y^m \rangle)$, i.e.,

$$\mathrm{Cap}(z_0^m, \theta) \triangleq \left\{ z^n \in \mathbb{S}^{m-1} : \angle(z_0^m, z^m) \leq \theta \right\}.$$

We will say that an arbitrary set $A \subseteq \mathbb{S}^{m-1}$ has an effective angle θ if $P_{Y^m}(A) = P_{Y^m}(C)$, where $C = \mathrm{Cap}(z_0^m, \theta)$ for some arbitrary $z_0^m \in \mathbb{S}^{m-1}$.

Proposition 3.2: Let $A \subseteq \mathbb{S}^{m-1}$ be an arbitrary set with effective angle θ. Then for any $\omega > \pi/2 - \theta$,

$$P_{Y^m}(A_\omega) \to 1 \text{ as } m \to \infty, \quad (19)$$

where A_ω is the ω-neighborhood of A defined as

$$A_\omega \triangleq \{x^m \in \mathbb{S}^{m-1} : \min_{z^m \in A} \angle(z^m, x^m) \leq \omega\}.$$

This results follows from the strengthening of Talagrand's inequality in (8). The bound in (8) allows to control the Wasserstein distance in terms of two separate parameters of the distribution P_Z, namely its second moment and its entropy. In the argument described around (18), this allows us to control both the measure of the typical set of P_Z and the radius of the sphere on which this set concentrates as $m \to \infty$, leading to the blowing-up lemma on the sphere, which can be shown to in turn imply Prop. 3.1. It is easy to see that when A is a spherical cap with angle θ, its blowup $A_{\frac{\pi}{2}-\theta+\epsilon}$ is a cap (slightly bigger than a halfsphere) whose probability approaches 1 in high dimensions. Therefore, when A is a spherical cap of angle θ, $\omega = \pi/2 - \theta + \epsilon$ is precisely the blowup angle needed for A_ω to approach probability 1. Prop. 3.2 asserts that the same blowup angle is sufficient for any other set A of the same measure, therefore effectively identifying the spherical cap as the extremal set for minimizing the measure of the neighborhood.

C. A New Measure Concentration Result on the Sphere

Perhaps even more interestingly, the transportation inequality (7) for information constrained OT leads to a new concentration of measure result on the sphere, which recovers Prop. 3.2 as a special case. This new result was recently proved in [8], [13] by using Riesz' rearrangement inequality [14] and can be stated as follows:

Proposition 3.3: Let $A \subseteq \mathbb{S}^{m-1}$ be an arbitrary set with effective angle θ and let Y^m be uniformly distributed on $\mathbb{S}^{m-1}$. For any $\omega \in (\pi/2 - \theta, \pi/2]$, let

$$V = P_{Y_m}(\mathrm{Cap}(z_0^m, \theta) \cap \mathrm{Cap}(y_0^m, \omega)),$$

where z_0^m, y_0^m are perpendicular to each other, i.e. $\angle(z_0^m, y_0^m) = \pi/2$. Then for any $\epsilon > 0$, we have

$$P_{Y^m}(\{y^m : P_{Y^m}(A \cap \mathrm{Cap}(y^m, \omega)) \geq e^{-m\epsilon} \cdot V\}) \to 1. \quad (20)$$

Note that an equivalent way to state the blowing-up lemma in Prop. 3.2 is the following: Let $A \subseteq \mathbb{S}^{m-1}$ be an arbitrary set with effective angle θ. Then for any $\omega > \pi/2 - \theta$

$$P_{Y^m}(\{y^m : P_{Y^m}(A \cap \mathrm{Cap}(y^m, \omega)) > 0\}) \to 1.$$

This is true because $P_{Y^m}(A \cap \mathrm{Cap}(y^m, \omega)) > 0$ iff $y^m \in A_\omega$. Prop. 3.3 extends Prop. 3.2 by providing a lower bound on the intersection measure of $A \cap \mathrm{Cap}(y^m, \omega)$, for $\omega > \pi/2 - \theta$. When A itself is a cap, (20) is straightforward and follows from the fact that Y^m w.h.p. concentrates around the equator at angle $\pi/2$ from the pole of A, and therefore the intersection of the two spherical caps is given by V w.h.p. Interestingly, Prop. 3.3 asserts that this intersection measure is w.h.p. lower bounded by V for an arbitrary A with the same measure. In other words, the spherical cap not only minimizes the measure of its neighborhood as captured by Prop. 3.2, but roughly speaking, also minimizes its intersection measure with the neighborhood of a randomly chosen point on the sphere.

IV. An Application to Network Information Theory

We next show that the new information constrained transportation inequality has an immediate application in network information theory, and in particular, can be used to recover the recent solution of a problem posed by Cover in 1987 [15] regarding the capacity of the relay channel.

To describe Cover's problem, consider a Gaussian primitive relay channel given by $Z = X + W_1$ and $Y = X + W_2$, where X denotes the source signal constrained to average power P, Z and Y denote the received signals of the relay and the destination respectively, and $W_1 \sim \mathcal{N}(0, N)$ and

$W_2 \sim \mathcal{N}(0,1)$ are Gaussian noises that are independent of each other and X. The relay channel is "primitive" in the sense that the relay is connected to the destination with an isolated bit pipe of capacity C_0. Let $C(C_0)$ denote the capacity of this channel as a function of C_0. What is the critical value of C_0 such that $C(C_0)$ first equals $C(\infty)$? This is a problem posed by Cover in *Open Problems in Communication and Computation*, Springer-Verlag, 1987 [15], which he calls "The Capacity of the Relay Channel".

This question was answered in a recent work [8], [9], which shows that $C(C_0)$ can not equal to $C(\infty)$ unless $C_0 = \infty$, regardless of the SNR of the Gaussian channels. This result follows as a corollary to a new upper bound developed in [8], [9] on the capacity of this channel, which builds on a strong data processing inequality (SDPI) for a certain Markov chain. The proof of this SDPI in [8], [9] is geometric and heavily relies on the new measure concentration result stated in Prop. 3.3. We next show that the transportation inequality we develop in the current paper can be used to directly establish this SDPI without going through Prop. 3.3, thereby significantly simplifying the proof in [8], [9]. We now state the SDPI and briefly illustrate how it leads to a new upper bound on the relay channel. We then prove it by using a conditional version of our transportation inequality (12).

A. A Strong Data Processing Inequality

Consider a long Markov chain

$$Y^n - X^n - Z^n - U_n, \tag{21}$$

with $Z^n = X^n + W_1^n$ and $Y^n = X^n + W_2^n$, where $\mathrm{E}[\|X^n\|^2] = nP$, $W_1^n \sim \mathcal{N}(0, NI_n)$, $W_2^n \sim \mathcal{N}(0, I_n)$, and X^n, W_1^n, W_2^n are mutually independent. For this long Markov chain, the following SDPI was established in [8], [9] and is the key step in resolving Cover's problem.

Proposition 4.1: For the Markov chain described in (21), if $I(Z^n;U_n|Y^n) \le nC_0$, then $I(X^n;U_n|Y^n)$ is upper bounded by:

$$\max_{C'\in[0,C_0]} \min_{R>0} \frac{n}{2} \ln \frac{P\left(N+1-2e^{-C'}\sqrt{N(1-e^{-2R})}\right) + N\left(1-e^{-2C'}(1-e^{-2R})\right)}{(P+1)Ne^{-2R}}.$$

Prop. 4.1 allows us to derive a new upper bound on the relay channel. In particular, if we use U_n to denote the relay's transmission over the bit pipe, then it is easy to see that $Y^n - X^n - Z^n - U_n$ for the relay channel satisfies the conditions of the Markov chain described in (21), and $I(Z^n;U_n|Y^n) \le nC_0$. Therefore, by Fano's inequality and Prop. 4.1 we can bound $C(C_0)$ by

$$\begin{aligned} C(C_0) &\le I(X^n;U_n,Y^n) + n\epsilon \\ &= I(X^n;Y^n) + I(X^n;U_n|Y^n) + n\epsilon \\ &\le \max_{C'\in[0,C_0]} \min_{R>0} \frac{n}{2} \ln \frac{P(N+1-2e^{-C'}\sqrt{N(1-e^{-2R})}) + N(1-e^{-2C'}(1-e^{-2R}))}{Ne^{-2R}} + n\epsilon. \end{aligned}$$

This bound turns out to be tight enough for resolving Cover's problem.

B. Proof via Transportation Inequality

To prove Prop. 4.1, we need the following lemma, which is a conditional version of our transportation inequality (12). The proof of this lemma is based on (12) and Jensen's inequality, and is omitted in the current paper.

Lemma 4.1: For the Markov chain (21), if $I(Z^n;U_n|X^n) = nC'$ for some $C' \ge 0$, then for any $R > 0$ there exists a random vector $\bar{Z}^n$ such that:

1) $P_{X^n,\bar{Z}^n,U_n} = P_{X^n,Z^n,U_n}$;
2) $\mathrm{E}[\bar{Z}^n \cdot Y^n] \ge n(P + \sqrt{N(1-e^{-2R})}e^{-C'})$;
3) $I(\bar{Z}^n;Y^n|X^n,U_n) \le nR$.

We now use Lemma 4.1 to prove Prop. 4.1. With $\bar{Z}^n$ coupled with Y^n, X^n, Z^n, U_n so as to satisfy the properties in Lemma 4.1, we have

$$\begin{aligned} &I(X^n;U_n|Y^n) \\ &= I(\bar{Z}^n;U_n|Y^n) + I(X^n;U_n|Y^n,\bar{Z}^n) - I(\bar{Z}^n;U_n|Y^n,X^n) \\ &= I(\bar{Z}^n;U_n|Y^n) + h(U_n|Y^n,\bar{Z}^n) - h(U_n|Y^n,X^n) \\ &\le I(\bar{Z}^n;U_n|Y^n) + h(U_n|\bar{Z}^n) - h(U_n|X^n) \\ &= I(\bar{Z}^n;U_n|Y^n) - I(Z^n;U_n|X^n) \qquad (22) \\ &= h(\bar{Z}^n|Y^n) - h(\bar{Z}^n|Y^n,U_n) - I(Z^n;U_n|X^n) \qquad (23) \end{aligned}$$

where (22) follows from 1) of Lemma 4.1, i.e. $P_{X^n,\bar{Z}^n,U_n} = P_{X^n,Z^n,U_n}$. Let the third term in (23) be

$$I(Z^n;U_n|X^n) = nC'. \tag{24}$$

Then we have $C' \in [0, C_0]$ because

$$I(Z^n;U_n|X^n) \le I(Z^n;U_n|Y^n) \le nC_0.$$

We now bound the first two terms in (23) respectively.

To bound the first term in (23), note that for any $R > 0$,

$$\begin{aligned} h(\bar{Z}^n|Y^n) &= h\left(\bar{Z}^n - \frac{\mathrm{E}[\bar{Z}^n \cdot Y^n]}{\mathrm{E}[\|Y^n\|^2]}Y^n \Big| Y^n\right) \\ &\le h\left(\bar{Z}^n - \frac{\mathrm{E}[\bar{Z}^n \cdot Y^n]}{\mathrm{E}[\|Y^n\|^2]}Y^n\right) \\ &\le \frac{n}{2}\ln\frac{2\pi e}{n}\mathrm{E}\left[\left\|\bar{Z}^n - \frac{\mathrm{E}[\bar{Z}^n \cdot Y^n]}{\mathrm{E}[\|Y^n\|^2]}Y^n\right\|^2\right] \\ &= \frac{n}{2}\ln\frac{2\pi e}{n}\left(\mathrm{E}[\|\bar{Z}^n\|^2] - \frac{\mathrm{E}[\bar{Z}^n \cdot Y^n]^2}{\mathrm{E}[\|Y^n\|^2]}\right) \\ &\le \frac{n}{2}\ln 2\pi e \frac{P(N+1-2e^{-C'}\sqrt{N(1-e^{-2R})}) + N(1-e^{-2C'}(1-e^{-2R}))}{P+1} \end{aligned} \tag{25}$$

where in the last step we have used 2) of Lemma 4.1.

To bound the second term in (23), we have for any $R > 0$,

$$\begin{aligned} h(\bar{Z}^n|Y^n,U_n) &\ge h(\bar{Z}^n|Y^n,U_n,X^n) \\ &= h(\bar{Z}^n|U_n,X^n) - I(\bar{Z}^n;Y^n|U_n,X^n) \\ &= h(Z^n|X^n) - I(Z^n;U_n|X^n) - I(\bar{Z}^n;Y^n|U_n,X^n) \\ &\ge \frac{n}{2}\ln 2\pi eN - nC' - nR \\ &= \frac{n}{2}\ln 2\pi N e^{1-2(C'+R)} \end{aligned} \tag{26}$$

where the second inequality follows from 3) of Lemma 4.1. Combining (23)–(26), we have proved Prop. 4.1.

REFERENCES

[1] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[2] L. V. Kantorovich. On translation of mass (in russian), c r. In *Doklady. Acad. Sci. USSR*, volume 37, pages 199–201, 1942.

[3] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6:587–600, May 1996.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[5] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[6] K. Marton. A simple proof of the blowing-up lemma (corresp.). *IEEE Transactions on Information Theory*, 32(3):445–446, 1986.

[7] K. Marton. Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

[8] X. Wu, L. P. Barnes, and A. Özgür. The capacity of the relay channel: Solution to cover's problem in the gaussian case. *IEEE Transactions on Information Theory*, 65(1):255–275, 2019.

[9] X. Wu, L. P. Barnes, and A. Özgür. The geometry of the relay channel. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2233–2237. IEEE, 2017.

[10] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

[11] D. Bakry, F. Bolley, and I. Gentil. Dimension dependent hypercontractivity for gaussian kernels. *Probability Theory and Related Fields*, 154(3-4):845–874, 2012.

[12] Olivier Rioul and Max H. M. Costa. On some almost properties. In *2016 Information Theory and Applications Workshop (ITA)*, pages 1–5, La Jolla, CA, USA, January 2016. IEEE.

[13] Leighton Pate Barnes, Ayfer Ozgur, and Xiugang Wu. An isoperimetric result on high-dimensional spheres. *arXiv preprint arXiv:1811.10533*, 2018.

[14] A. Baernstein II and B. A. Taylor. Spherical rearrangements, subharmonic functions, and $*$-functions in n-space. *Duke Mathematical Journal*, 43(2):245–268, 1976.

[15] T. M. Cover. The capacity of the relay channel. In *Open Problems in Communication and Computation*, pages 72–73. Springer, 1987.