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

Yikun Bai, Xiugang Wu, *Member, IEEE*, and Ayfer Özgür, *Member, IEEE*

***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 new inequality to network information theory. We show that it can be used to recover almost immediately a recent solution to a long-standing open problem posed by Cover regarding the capacity of the relay channel.**

***Index Terms*— Optimal transport (OT), information constraint, transportation inequality, isoperimetric inequality, concentration of measure, network information theory, relay channel.**

I. Introduction

THE optimal transport (OT) theory, pioneered by Monge [2] and Kantorovich [3], 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 Borel 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 optimal coupling in $\Pi(P_Z,P_Y)$ that minimizes the expected cost:

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

A special case of particular interest is when both $\mathcal{Z}$ and $\mathcal{Y}$ are the Euclidian space and the cost function is given by the Euclidian distance. For simplicity, let us for now consider the one-dimensional case where $\mathcal{Z}=\mathcal{Y}=\mathbb{R}$ and $c(z,y)=|z-y|^p$; the generalization to arbitrary dimensions will be formalized and discussed in the subsequent sections. In this case, the quantity

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

defines a metric 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 [4] 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 propose to 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:

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

There are several reasons for us to study this extension of the classical OT problem, which will become clear in the sequel. For now, note that when the infimum in (1) is achieved by a deterministic mapping between Z and Y, the mutual information $I_P(Z;Y)$ will be maximal and can be potentially unbounded. For example, according to Brenier's theorem [5], this is known to be the case in (2) when $p=2$ and P_Z or

Manuscript received 24 August 2020; revised 21 July 2021; accepted 26 October 2021. Date of publication 16 January 2023; date of current version 17 March 2023. This work was supported by NSF under Award CIF-1704624, Award CIF-2213223, and Award NeTS-1817205. An earlier version of this paper was presented in part at the 2020 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT44484.2020.9174478]. *(Corresponding author: Xiugang Wu.)*

Yikun Bai was with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA. He is now with the Department of Computer Science, Vanderbilt University, Nashville, TN 37212 USA (e-mail: yikun.bai@vanderbilt.edu).

Xiugang Wu is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: xwu@udel.edu).

Ayfer Özgür is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: aozgur@stanford.edu).

Communicated by M. Raginsky, Associate Editor for Probability and Statistics.

Digital Object Identifier 10.1109/TIT.2023.3237073

P_Y are absolutely continuous with respect to the Lebesgue measure. The mutual information constraint in (4) can be viewed as enforcing a certain amount of randomization in the mapping between Z and Y. On a related note, formulation (4) has also been considered in the context of rate-distortion theory when one is interested in the distortion-rate function for fixed source and output distributions P_Z and P_Y; see [6], where the authors call such problems 'output-constrained distortion-rate function.'

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

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

The problem (5) generally appears under the name entropy regularized OT. In the machine learning literature, the 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 classical OT in (1), which motivates the use of (5) instead of (1) as a discrepancy measure between probability measures [7]. For certain inference tasks, (5) also appears to be a more suitable discrepancy measure than (1), leading to superior empirical performance [8]. Moreover, it is also shown in [9] and [10] that (5) can be estimated with much fewer samples as compared to (1). In contrast to these works which focus on the computational and statistical aspects of (5), our interest in this paper mainly lies 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:

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

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 $P_Z \ll P_Y$:

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

This new transportation inequality captures the trade-off between information constraint and transportation cost, and is tight when P_Z is Gaussian. It can be regarded as a generalization and sharpening of Talagrand's inequality in (3). Note that when we take $R\to\infty$ in (7), we get the following bound on the unconstrained Wasserstein distance:

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

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).

Since the pioneering work of Marton [11], [12], it has been known that Talagrand's transportation inequality 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. Do the new transportation inequalities in (7) and (8) also have natural geometric counterparts? 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 the Gaussian space. In other words, the strengthening of Talagrand's inequality in (8) captures a stronger isoperimetric phenomenon, 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 [13], [14]. Different from the standard isoperimetric inequality on the sphere where one is interested in the extremal set that minimizes the measure of its neighborhood among all sets of equal measure, 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 information constrained transportation inequality (7) to network information theory. In particular, we show that it can be used to understand and simplify the recent solution of a long-standing open problem on communication over the three-node relay channel. Specifically, this problem, "The Capacity of the Relay Channel", was posed by Cover in the book *Open Problems in Communication and Computation*, Springer-Verlag, 1987 [15]. The recent works [14], [16] solved this problem in the canonical Gaussian case by developing a new converse for the relay channel.[1] The proof in [14] and [16] is geometric: the communication problem is recast as a problem about the geometry of typical sets in high-dimensions, and then solved using the new isoperimetric result on the sphere mentioned above. The new transportation inequality (7) allows us to recover the same result almost immediately, which also enables an interpretation of the previous geometric proof in terms of auxiliary random variables.

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 [18]. 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

$$\mathbb{E}_{\omega\sim\nu}[d^p(\omega,\omega_0)] < \infty \tag{9}$$

holds for some (and hence all) $\omega_0\in\Omega$.

Definition 2.1 (Wasserstein Distance): Given $p\ge 1$, the Wasserstein distance of order p between any pair P_Z,

[1] See also [17] for the solution in the case of binary symmetric channels.

$P_Y \in \mathcal{P}_p(\Omega)$ is defined as

$$W_p(P_Z, P_Y) \triangleq \left\{ \inf_{P \in \Pi(P_Z, P_Y)} \mathbb{E}_P[d^p(Z, Y)] \right\}^{1/p} \tag{10}$$

where $\Pi(P_Z, P_Y)$ is the set of all probability measures on the product space $\Omega \times \Omega$ with marginals P_Z and P_Y.

Indeed, the function $W_p(P_Z, P_Y)$ of (P_Z, P_Y) in (10) satisfies all the metric axioms [19] and defines a metric on the space $\mathcal{P}_p(\Omega)$ of distributions. 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 (10) is given by the deterministic mapping

$$Z = F_Z^{-1} \circ F_Y(Y) \tag{11}$$

where F_Y is the cdf of P_Y, i.e. $F_Y(y) = P_Y(Y \leq y)$ and F_Z^{-1} is the quantile function of P_Z, i.e. $F_Z^{-1}(\alpha) = \inf\{z \in \mathbb{R} : F_Z(z) \geq \alpha\}$. Building on this optimal coupling and tensorization [18], 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 [4]): For two probability measures $P_{Z^n} \ll P_{Y^n}$ on $\mathbb{R}^n$ with P_{Y^n} being standard Gaussian $\mathcal{N}(0, I_n)$, we have

$$W_2^2(P_{Z^n}, P_{Y^n}) \leq 2D(P_{Z^n} \| P_{Y^n}), \tag{12}$$

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 $\mu \in \mathbb{R}^n$.

Note that Talagrand's transportation inequality connects the Wasserstein distance to another fundamental discrepancy measure between two probability measures, i.e. the KL divergence. This observation can be utilized to derive, via Marton's procedure [11], [12], the concentration of measure phenomenon in the Gaussian space—that is, for all subsets with a given measure in the Gaussian space, blowing up the set with a minimum needed radius will increase its probability to nearly 1. The rigorous Marton's argument for connecting transportation and concentration in this case will be briefly illustrated in Section III-A. Here, it is of interest to point out the intuition behind this connection from an information theoretic point of view: if we think of a subset in the Gaussian space as the typical set generated by P_{Z^n}, then the Wasserstein distance on the L.H.S. of (12) will translate to the minimum needed blowing-up radius and knowing $D(P_{Z^n} \| P_{Y^n})$ on the R.H.S. of (12) is equivalent to fixing the measure of the subset in the Gaussian space, and hence the transportation inequality (12) establishes a relationship between the minimum needed blowing-up radius for increasing the probability of the set to nearly 1 and the measure of the set.

A. Sharpening Talagrand's Transportation Inequality

Talagrand's transportation inequality can be sharpened to the following; see also [20] and [21] for related results.

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

$$W_2^2(P_{Z^n}, P_{Y^n}) \leq \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n}h(Z^n)}}, \tag{13}$$

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 $\sigma > 0$.

The above theorem provides an upper bound (13) on the Wasserstein distance between $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} \ll P_{Y^n}$ in terms of the second moment $\mathbb{E}[\|Z^n\|^2]$ and entropy $h(Z^n)$ of Z^n, instead of in terms of the KL divergence $D(P_{Z^n} \| P_{Y^n})$ as in Talagrand's transportation inequality. Note that given $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} \ll P_{Y^n}$, one can always use $\mathbb{E}[\|Z^n\|^2]$ and $h(Z^n)$ to synthesize $D(P_{Z^n} \| P_{Y^n})$ via

$$D(P_{Z^n} \| P_{Y^n}) = -h(Z^n) + \frac{n}{2} \ln 2\pi + \frac{n}{2} \mathbb{E}[\|Z^n\|^2]$$

but not vice versa. In other words, compared to knowing only $D(P_{Z^n} \| P_{Y^n})$, one has more information about P_{Z^n} knowing both $\mathbb{E}[\|Z^n\|^2]$ and $h(Z^n)$, and a natural question to ask now is whether one can better bound $W_2(P_{Z^n}, P_{Y^n})$ using $\mathbb{E}[\|Z^n\|^2]$ and $h(Z^n)$. To this end, we show in Appendix A that the transportation inequality (13) is indeed generally stronger than Talagrand's, i.e. R.H.S. of (13) $\leq$ R.H.S. of (12), for any $P_{Z^n} \ll P_{Y^n}$. Moreover, note that compared to Talagrand's transportation inequality, which is tight only when $P_{Z^n} = \mathcal{N}(\mu, I_n)$, the inequality (13) is tight for a wider class of P_{Z^n}, i.e. when P_{Z^n} is isotropic Gaussian.

Just like Talagrand's transportation inequality implies concentration of measure in the Gaussian space, the transportation inequality (13) also has its own geometric interpretation. In particular, we will show in Section III-B that (13) can be used to prove concentration of measure on the sphere, instead of in the Gaussian space—that is, for all subsets with a given measure on the sphere, blowing up the set with a minimum needed angle will increase its probability to nearly 1. Note that the concentration on the sphere is known to be stronger than that in the Gaussian space, and this should not be surprising as alluded by the fact that (13) implies Talagrand's transportation inequality. An intuitive explanation for why (13) corresponds to the concentration on the sphere is because if we think of the subset as the typical set generated by P_{Z^n}, then knowing both the second moment $\mathbb{E}[\|Z^n\|^2]$ and entropy $h(Z^n)$ of Z^n amounts to restricting the subsets to those on the sphere with a fixed measure.

B. Information Constrained OT

We next focus on bounding the information constrained OT.

Definition 2.2 (Information Constrained Wasserstein Divergence): Given $p \geq 1$, the Wasserstein divergence of order p between any pair $P_Z, P_Y \in \mathcal{P}_p(\Omega)$ subject to information constraint R is defined as

$$W_p(P_Z, P_Y; R) \triangleq \left\{ \inf_{\substack{P \in \Pi(P_Z, P_Y): \\ I_P(Z;Y) \leq R}} \mathbb{E}_P[d^p(Z, Y)] \right\}^{1/p}. \tag{14}$$

It can be verified that the function $W_p(P_Z, P_Y; R)$ of (P_Z, P_Y) in (14) is nonnegative, symmetric in (P_Z, P_Y), and satisfies the triangle inequality (see [7]). However, it is not a metric (and hence is called a 'divergence' instead of a 'distance') because it violates the coincidence axiom, i.e., $W_p(P_Z, P_Y; R)$ in general is not equal to zero when

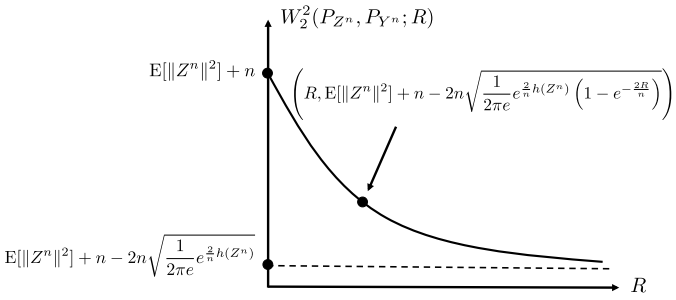

Fig. 1. Information constraint-Wasserstein divergence tradeoff.

$P_Z = P_Y$. 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 it.

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

$$W_2^2(P_{Z^n}, P_{Y^n}; R) \leq \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}\left(1 - e^{-\frac{2R}{n}}\right)}. \quad (15)$$

The above theorem characterizes a trade-off between the information constraint and the Wasserstein divergence, as depicted in Fig. 1. This includes Theorem 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 divergence simply reduces to $\mathbb{E}[\|Z^n\|^2] + n$. In Appendix B, we show that the new transportation inequality (15) is tight when P_{Z^n} is isotropic Gaussian; that is, when $P_{Z^n} = \mathcal{N}(\mu, \sigma^2 I_n)$ for some μ and σ^2, the inequality in (15) is achieved with equality. Therefore, the trade-off characterized in Theorem 2.2 is indeed tight when P_{Z_n} is isotropic Gaussian.

Geometrically, the information constrained transportation inequality (15) turns out to capture a new concentration phenomenon on the sphere that is recently proved in [14]. This new concentration result extends the classical concentration result on the sphere mentioned in Section II-A, and provides a lower bound on the measure of the intersection between a subset with a given measure on the sphere and the neighborhood of a randomly chosen point on the sphere. Intuitively, (15) allows one to control the intersection measure because the information constraint $I_P(Z^n; Y^n) \leq R$ is equivalent to the conditional entropy constraint $H_P(Z^n|Y^n) \geq H(Z^n) - R$, and the latter can be thought of as putting a lower bound on the measure of the conditional typical set of Z^n given a typical sequence of Y^n, which is further a lower bound on the intersection measure between the typical set of Z^n and the neighborhood of a randomly chosen point on the sphere given $P_{Y^n} = \mathcal{N}(0, I_n)$. Rigorously, to derive this new concentration result from the transportation inequality, we need a more technical version of (15), namely an information density constrained transportation inequality. We will introduce this information density constrained transportation inequality in the sequel and use it to prove the new concentration result in Section III-C.

C. Conditional Transportation Inequality

Both Theorems 2.1 and 2.2 have their conditional versions. We start by defining the conditional Wasserstein distance and the conditional information constrained Wasserstein divergence.

Definition 2.3 (Conditional Wasserstein Distance): Fix a probability measure P_T and two conditional probability measures $P_{Z|T}$ and $P_{Y|T}$ with $P_{Z|T=t}, P_{Y|T=t} \in \mathcal{P}_p(\Omega)$ for any t. Given $p \geq 1$, the conditional Wasserstein distance of order p between $P_{Z|T}, P_{Y|T}$ given P_T is defined as

$$W_p(P_{Z|T}, P_{Y|T}|P_T) \triangleq \left\{\inf_{P \in \Pi(P_{Z|T}, P_{Y|T}|P_T)} \mathbb{E}_P[d^p(Z, Y)]\right\}^{1/p} \quad (16)$$

where

$$\Pi(P_{Z|T}, P_{Y|T}|P_T) \triangleq \{P_{\bar{Z},\bar{Y}|T} \cdot P_T : P_{\bar{Z}|T} = P_{Z|T}, P_{\bar{Y}|T} = P_{Y|T}\}. \quad (17)$$

Theorem 2.3: For any probability measure P_T and conditional probability measures $P_{Z^n|T}$ and $P_{Y^n|T}$ such that for any t, $P_{Y^n|T=t} = P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n|T=t} \ll P_{Y^n}$, we have

$$W_2^2(P_{Z^n|T}, P_{Y^n|T}|P_T) \leq \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|T)}}. \quad (18)$$

Definition 2.4 (Conditional Information Constrained Wasserstein Divergence): Fix a probability measure P_T and two conditional probability measures $P_{Z|T}$ and $P_{Y|T}$ with $P_{Z|T=t}, P_{Y|T=t} \in \mathcal{P}_p(\Omega)$ for any t. Given $p \geq 1$, the conditional Wasserstein divergence of order p between $P_{Z|T}, P_{Y|T}$ given P_T subject to information constraint R is defined as

$$W_p(P_{Z|T}, P_{Y|T}|P_T; R) \triangleq \left\{\inf_{\substack{P \in \Pi(P_{Z|T}, P_{Y|T}|P_T),\\ I_P(Z;Y|T) \leq R}} \mathbb{E}_P[d^p(Z, Y)]\right\}^{1/p}. \quad (19)$$

Theorem 2.4: For any probability measure P_T and conditional probability measures $P_{Z^n|T}$ and $P_{Y^n|T}$ such that for any t, $P_{Y^n|T=t} = P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n|T=t} \ll P_{Y^n}$, we have

$$W_2^2(P_{Z^n|T}, P_{Y^n|T}|P_T; R) \leq \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|T)}\left(1 - e^{-\frac{2R}{n}}\right)}. \quad (20)$$

D. Information Density Constrained OT

We now introduce an OT setup with information density constraint, and present a transportation inequality for this new setup. As we will see, the information density constraint is more stringent than the information constraint, and therefore our previous transportation inequality in Theorem 2.2 can be viewed as a special case of this new inequality that we are going to present. This new inequality will be used to prove a new concentration of measure result on the sphere, which

has not been known before the recent work [13], [14]; see Proposition 3.3 and its proof in the next section.

Recall that for a given joint distribution $P \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y})$ with marginals P_Z and P_Y, the information density function $i_P(z;y)$ is defined as

$$i_P(z;y) = \ln \frac{dP}{dP_Z \otimes P_Y}(z,y),$$

whose expectation gives rise to the mutual information $I_P(Z;Y)$, i.e.,

$$I_P(Z;Y) = \mathbb{E}_P[i_P(Z;Y)].$$

We say that a distribution P satisfies (R,τ,δ)-information density constraint for some $R \geq 0$ and $\tau, \delta > 0$, if the following two conditions hold:

1) the expectation of the information density, i.e. the mutual information, is upper bounded by R,

$$I_P(Z;Y) \leq R;$$

2) with probability at least $1-\delta$, the deviation between the information density and mutual information is upper bounded by τ,

$$\mathbb{P}_{(Z,Y)\sim P}(|i_P(Z;Y) - I_P(Z;Y)| \leq \tau) \geq 1-\delta.$$

Definition 2.5: (Information Density Constrained Wasserstein Divergence): Given $p \geq 1$, the Wasserstein divergence of order p between any pair $P_Z, P_Y \in \mathcal{P}_p(\Omega)$ subject to (R,τ,δ)-information density constraint is defined as

$$W_p(P_Z,P_Y;R,\tau,\delta) \triangleq \left\{ \inf_{\substack{P\in\Pi(P_Z,P_Y): I_P(Z;Y)\leq R,\\ \mathbb{P}_{(Z,Y)\sim P}(|i_P(Z;Y)-I_P(Z;Y)|\leq\tau)\geq 1-\delta}} \mathbb{E}_P[d^p(Z,Y)] \right\}^{1/p}. \tag{21}$$

Compared to the information constrained case, the definition of the information density constrained Wasserstein divergence in (21) involves an additional constraint $\mathbb{P}_{(Z,Y)\sim P}(|i_P(Z;Y) - I_P(Z;Y)| \leq \tau) \geq 1-\delta$ in the infimization and therefore given arbitrary P_Z, P_Y and R we have

$$W_p(P_Z,P_Y;R) \leq W_p(P_Z,P_Y;R,\tau,\delta)$$

for any $\tau, \delta > 0$. As in the information constrained case, the quantity $W_p(P_Z,P_Y;R,\tau,\delta)$ is not a metric because $W_p(P_Z,P_Y;R,\tau,\delta)$ in general is not equal to zero when $P_Z = P_Y$.

For this OT setup with information density constraint, we have the following bound when $\Omega = \mathbb{R}^n$ and $d(z^n,y^n) = \|z^n - y^n\|_2$.

Theorem 2.5: For $P_{Y^n} = \mathcal{N}(0,I_n)$ and $P_{Z^n} \ll P_{Y^n}$, we have that for any $R, \tau \geq 0$

$$W_2^2(P_{Z^n},P_{Y^n};R,\tau,6n/\tau^2) \leq \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n}h(Z^n)}\left(1-e^{-\frac{2R}{n}}\right)} \tag{22}$$

It is easy to see that the above theorem includes Theorem 2.2 as a special case by noting that

$$W_2^2(P_{Z^n},P_{Y^n};R) \leq W_2^2(P_{Z^n},P_{Y^n};R,\tau,6n/\tau^2)$$

for any $R, \tau \geq 0$.

E. Proofs of New Transportation Inequalities

In this subsection, we provide the proofs of Theorems 2.1–2.5. Recall that Theorems 2.1, 2.2 and 2.5 are unconditional transportation inequalities, while Theorems 2.3 and 2.4 are the conditional versions. In particular, Theorem 2.1 follows from Theorem 2.2, which in turn follows from Theorem 2.5 as a special case. Thus, in the following we first focus on proving Theorem 2.5 to establish all the unconditional transportation inequalities stated in the paper. Then we show how to obtain the conditional versions, in particular Theorems 2.3 and 2.4; for this, it suffices to show how to extend Theorem 2.2 to Theorem 2.4.

Proof of Theorem 2.5: To show Theorem 2.5, it suffices to construct a coupling P of P_{Z^n} and P_{Y^n} such that the $(R,\tau,6n/\tau^2)$-information density constraint is satisfied, i.e.,

$$I_P(Z^n;Y^n) \leq R$$

and

$$\mathbb{P}_{(Z^n,Y^n)\sim P}(|i_P(Z^n;Y^n) - I_P(Z^n;Y^n)| \leq \tau) \geq 1 - 6n/\tau^2,$$

and simultaneously $\mathbb{E}_P[\|Z^n - Y^n\|^2]$ is upper bounded by the R.H.S. of (22). For this, let

$$Y^n = \sqrt{1-e^{-\frac{2R}{n}}}\, Y_1^n + e^{-\frac{R}{n}} Y_2^n,$$

where $Y_1^n, Y_2^n \sim \mathcal{N}(0,I_n)$ are independent of each other, and let Z^n satisfy

$$Z^n = g(Y_1^n)$$

for some $g : \mathbb{R}^n \to \mathbb{R}^n$ that pushes $P_{Y_1^n} = \mathcal{N}(0,I_n)$ forward to P_{Z^n}, where g is a differentiable one-to-one mapping whose Jacobian matrix J_g has only nonnegative eigenvalues. Note that such a mapping g always exists provided that $P_{Y_1^n} = \mathcal{N}(0,I_n)$ is absolute continuous with respect to the Lebesgue measure and $P_{Z^n} \ll P_{Y_1^n}$, and examples include the Brenier mapping [5] and the Knothe-Rosenblatt mapping [19]; see also Lemma 1 of [21].

It is easy to verify that the joint distribution P of (Z^n,Y^n) defined by the above is indeed a coupling of P_{Z^n} and P_{Y^n}. (In fact, this coupling is closely related to the concepts of Ornstein–Uhlenbeck semi-group and Ornstein–Uhlenbeck process [22] as we illustrate in Appendix C.[2]) To see that this coupling satisfies the information density constraint, first note that

$$\begin{aligned} I_P(Z^n;Y^n) &= h(Y^n) - h(\sqrt{1-e^{-\frac{2R}{n}}}\, Y_1^n + e^{-\frac{R}{n}} Y_2^n | Z^n) \\ &= h(Y^n) - h(e^{-\frac{R}{n}} Y_2^n | Z^n) \end{aligned} \tag{23}$$

[2]On a related note, see [23] for the application of reverse hypercontractivity results for the Ornstein–Uhlenbeck process to the relay channel problem, though both the proof techniques and the results in [23] are different (and significantly weaker) than those to be presented in Section IV of the current paper.

$$\begin{aligned} &= h(Y^n) - h(e^{-\frac{R}{n}} Y_2^n) \qquad (24)\\ &= h(Y^n) - n\ln(e^{-\frac{R}{n}}) - h(Y_2^n)\\ &= R \end{aligned}$$

where (23) holds because g is a one-to-one mapping and thus Y_1^n is determined given Z^n, and (24) follows from the independence between Y_2^n and Z^n. Also we have

$$\begin{aligned} &\mathbb{P}(|i_P(Z^n;Y^n) - I_P(Z^n;Y^n)| \le \tau)\\ &= \mathbb{P}\left(\left|\ln\left(\frac{f_{Y^n|Z^n}(Y^n|Z^n)}{f_{Y^n}(Y^n)}\right) - R\right| \le \tau\right)\\ &= \mathbb{P}\left(\left|\ln\left(\frac{f_{e^{-R/n}Y_2^n}(e^{-R/n}Y_2^n)}{f_{Y^n}(Y^n)}\right) - R\right| \le \tau\right)\\ &= \mathbb{P}(|\|Y^n\|^2 - \|Y_2^n\|^2| \le 2\tau)\\ &\ge \mathbb{P}(|\|Y^n\|^2 - n| \le \tau, |\|Y_2^n\|^2 - n| \le \tau)\\ &\ge 1 - (\mathbb{P}(|\|Y^n\|^2 - n| \ge \tau) + \mathbb{P}(|\|Y_2^n\|^2 - n| \ge \tau))\\ &\ge 1 - \frac{6n}{\tau^2} \qquad (25) \end{aligned}$$

where (25) holds by Chebyshev's inequality.

Now it remains to show $\mathbb{E}_P[\|Z^n - Y^n\|^2] \le$ R.H.S of (22). For this, we will lower bound $\mathbb{E}_P[Z^n \cdot Y^n]$ in the sequel. In particular, letting g_i denote the ith coordinate of g, we have

$$\begin{aligned} \mathbb{E}_P[Z^n \cdot Y_1^n] &= \sum_{i=1}^n \mathbb{E}_P\left[\frac{\partial g_i}{\partial y_{1i}}(Y_1^n)\right] \qquad (26)\\ &= \mathbb{E}_P[\text{trace}(J_g(Y_1^n))]\\ &\ge \mathbb{E}_P[n(\det(J_g(Y_1^n)))^{1/n}] \qquad (27)\\ &= n\mathbb{E}_P[e^{\ln(\det(J_g(Y_1^n)))^{1/n}}] \qquad (28)\\ &\ge ne^{\frac{1}{n}\mathbb{E}_P \ln[\det(J_g(Y_1^n))]} \qquad (29)\\ &= ne^{\frac{1}{n}(h(Z^n) - h(Y_1^n))} \qquad (30)\\ &= n\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}} \end{aligned}$$

where (26) follows from Stein's lemma for $P_{Y_1^n} = \mathcal{N}(0, I_n)$, which says that if $Y_1^n \sim \mathcal{N}(0, I_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then $\mathbb{E}[f(Y_1^n)Y_{1i}] = \mathbb{E}[\frac{\partial}{\partial y_i} f(Y_1^n)]$; (27) holds by the fact that for any matrix A whose eigenvalues are all nonnegative, $\frac{1}{n}\text{trace}(A) \ge (\det(A))^{1/n}$; (28) follows from the nonnegativity of $\det(J_g(Y_1^n))$; (29) is due to Jensen's inequality; and (30) holds because $Z^n = g(Y_1^n)$ and therefore

$$f_{Z^n}(g(y_1^n))\det(J_g(y_1^n)) = f_{Y_1^n}(y_1^n), \forall y_1^n.$$

Therefore, $\mathbb{E}_P[Z^n \cdot Y^n]$ is lower bounded by

$$\begin{aligned} \mathbb{E}_P[Z^n \cdot Y^n] &= \sqrt{1 - e^{-\frac{2R}{n}}}\mathbb{E}_P[Z^n \cdot Y_1^n]\\ &\ge n\sqrt{1 - e^{-\frac{2R}{n}}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}}, \end{aligned}$$

and hence

$$\begin{aligned} &\mathbb{E}_P[\|Z^n - Y^n\|^2]\\ &= \mathbb{E}[\|Z^n\|^2] + n - 2\mathbb{E}_P[Z^n \cdot Y^n]\\ &\le \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}\left(1 - e^{-\frac{2R}{n}}\right)}. \end{aligned}$$

This completes the proof of Theorem 2.5. ∎

We now show how to obtain Theorem 2.4 based on Theorem 2.2.

Proof of Theorem 2.4: By Theorem 2.2, there exists some $P_{\bar{Z}^n,\bar{Y}^n|T}$ such that for any t

$$P_{\bar{Y}^n|T=t} = \mathcal{N}(0, I_n),\ P_{\bar{Z}^n|T=t} = P_{Z^n|T=t}, \qquad (31)$$

$$I(\bar{Z}^n;\bar{Y}^n|T=t) \le R, \qquad (32)$$

$$\text{and}\ \ \mathbb{E}[\bar{Z}^n \cdot \bar{Y}^n|T=t] \ge n\sqrt{1 - e^{-\frac{2R}{n}}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|T=t)}}. \qquad (33)$$

Since (31) holds for any t, we have $P_{\bar{Y}^n|T} = \mathcal{N}(0, I_n)$ and $P_{\bar{Z}^n|T} = P_{Z^n|T}$, and hence

$$P_{\bar{Z}^n,\bar{Y}^n|T} \in \Pi(P_{Z^n|T}, P_{Z^n|T}|P_T). \qquad (34)$$

From (32), we get

$$\begin{aligned} I(\bar{Z}^n;\bar{Y}^n|T) &= \mathbb{E}_{S\sim P_T}[I(\bar{Z}^n;\bar{Y}^n|T=S)]\\ &\le \mathbb{E}_{S\sim P_T}[R]\\ &= R. \qquad (35) \end{aligned}$$

Moreover, using (33) we can lower bound $\mathbb{E}[\bar{Z}^n \cdot \bar{Y}^n]$ by

$$\begin{aligned} \mathbb{E}[\bar{Z}^n \cdot \bar{Y}^n] &= \mathbb{E}_{S\sim P_T}[\mathbb{E}[\bar{Z}^n \cdot \bar{Y}^n|T=S]]\\ &\ge \mathbb{E}_{S\sim P_T}\left[n\sqrt{1 - e^{\frac{2R}{n}}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|T=S)}}\right]\\ &\ge n\sqrt{1 - e^{\frac{2R}{n}}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}\mathbb{E}_{S\sim P_T}[h(Z^n|T=S)]}} \qquad (36)\\ &= n\sqrt{1 - e^{\frac{2R}{n}}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|T)}}, \end{aligned}$$

where (36) follows from Jensen's inequality, and therefore $\mathbb{E}[\|\bar{Z}^n - \bar{Y}^n\|^2]$ can be upper bounded by

$$\begin{aligned} &\mathbb{E}[\|\bar{Z}^n - \bar{Y}^n\|^2]\\ &= \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{1 - e^{\frac{2R}{n}}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|T)}}. \qquad (37) \end{aligned}$$

Combining (34), (35) and (37) completes the proof of Theorem 2.4. ∎

III. Geometry: Concentration and Isoperimetry

Since the pioneering work of Marton [11], [12], it has been known that transportation cost inequalities can be used to derive concentration of measure, an inherently geometric phenomenon tightly coupled with isoperimetric inequalities. For example, Talagrand's transportation inequality (12) can be shown to imply concentration of measure in the Gaussian space. In this section, we will discuss the geometric implications of the new transportation inequalities introduced and proved in the last section. In particular, we will show that the transportation inequalities stated in Theorem 2.1 and 2.5 can be used to prove the classical and new concentration results on the sphere. This leads to a more complete view on the interplay between transportation and concentration, as summarized in Fig. 2. We now begin the detailed discussion with the geometry of Talagrand's transportation inequality.

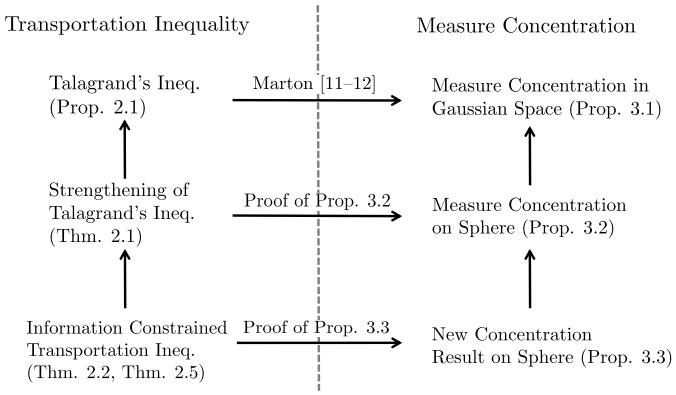

Fig. 2. Transportation and concentration.

A. Concentration and Isoperimetry in Gaussian Space

Consider a Gaussian space $(\mathbb{R}^n, \gamma)$, where $\gamma = \mathcal{N}(0, I_n)$ is the standard Gaussian measure on $\mathbb{R}^n$. For any $A \subseteq \mathbb{R}^n$ and $t > 0$, let A_t denote the t-blowup set of A:

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

The following concentration of measure result is generally known as the blowing-up lemma in Gaussian space [18].

Proposition 3.1: For any $A \subseteq \mathbb{R}^n$ with $\gamma_n(A) \geq e^{-na}$,

$$\gamma_n(A_t) \to 1 \text{ as } n \to \infty$$

when $t \geq \sqrt{2n(a+\epsilon)}$ 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. This lemma can be thought of as a consequence of the isoperimetric inequality in Gaussian space, which says that among all sets with equal Gaussian measure, a halfspace minimizes the measure of its t-blowup. Therefore, if we start with two sets A and H, where $\gamma(A) = \gamma(H)$ and H is a halfspace, then $\gamma(A_t) \geq \gamma(H_t)$ and hence it suffices to check that $\gamma(H_t) \to 1$, which follows from a simple calculation.

An alternative approach to proving the above blowing-up lemma, pioneered by Marton [11], [12], is through Talagrand's transportation inequality. A formal proof via this approach can be found in [18]. The key observation here is that for any measure ν and set A, $\nu(A)$ can be related to the KL divergence as

$$D(\nu_A \| \nu) = \ln \frac{1}{\nu(A)},$$

where ν_A is the conditional probability measure defined as $\nu_A(C) \triangleq \nu(C \cap A)/\nu(A)$ for any C. Together with the triangle inequality for the Wasserstein distance, this allows us to conclude that, for any $A, B \subseteq \mathbb{R}^n$,

$$\begin{aligned} W_2(\gamma_A, \gamma_B) &\leq W_2(\gamma_A, \gamma) + W_2(\gamma_B, \gamma) \\ &\leq \sqrt{2D(\gamma_A \| \gamma)} + \sqrt{2D(\gamma_B \| \gamma)} \\ &= \sqrt{2 \ln \frac{1}{\gamma(A)}} + \sqrt{2 \ln \frac{1}{\gamma(B)}}, \end{aligned}$$

where the second inequality follows from Talagrand's transportation inequality in (12). The proof of Proposition 3.1 follows by taking $B = A_t^c \triangleq \mathbb{R}^n \setminus A_t$ and noting that $W_2(\gamma_A, \gamma_{A_t^c}) \geq t$.

B. Concentration and Isoperimetry on the Sphere

We next show that the stronger transportation inequality (7) also has a natural geometric counterpart. In particular, it implies the following concentration result on the sphere: Consider a unit sphere $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ equipped the uniform probability measure μ on $\mathbb{S}^{n-1}$, denoted by $(\mathbb{S}^{n-1}, \mu)$, where

$$\mathbb{S}^{n-1} = \{z^n \in \mathbb{R}^n : \|z^n\| = 1\}.$$

Recall that a spherical cap with angle θ on $\mathbb{S}^{n-1}$ is defined as a ball on $\mathbb{S}^{n-1}$ in the geodesic metric (or simply the angle) $\angle(z^n, y^n) = \arccos(\langle z^n, y^n \rangle)$, i.e.,

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

Using the formula for the area of a spherical cap (see [14, Appendix C]), we can show that as $n \to \infty$

$$\mu(A)^{1/n} \to \sin\theta. \tag{38}$$

Proposition 3.2: Let $A \subseteq \mathbb{S}^{n-1}$ be an arbitrary set with $\mu(A) = \mu(\text{Cap}(z_0^n, \theta))$ for some arbitrary $z_0^n \in \mathbb{S}^{n-1}$ and $\theta \in (0, \pi/2]$. Then for any $\omega > \pi/2 - \theta$,

$$\mu(A_\omega) \to 1 \text{ as } n \to \infty, \tag{39}$$

where A_ω is the ω-blowup of A defined as

$$A_\omega \triangleq \{x^n \in \mathbb{S}^{n-1} : \angle(z^n, x^n) \leq \omega \text{ for some } z^n \in A\}.$$

As in the case of Gaussian measure concentration, the result in Proposition 3.2 is tightly related to the isoperimetric inequality on the sphere. It is easy to see that when A is a spherical cap with angle θ, its blowup $A_{\frac{\pi}{2}-\theta+\epsilon}$ is also 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. According to the isoperimetric inequality on the sphere [24], among all sets on the sphere with a given measure, the spherical cap is the extremal set for minimizing the measure of its blowup; therefore, the same blowup angle must be sufficient for any other set A with the same measure, and this is precisely what Proposition 3.2 asserts. We next show that Proposition 3.2 can be derived by properly combining the strengthening (13) of Talagrand's transportation inequality with an argument similar to Marton's procedure.

We next show that Proposition 3.2 can be derived by properly combining the strengthening (13) of Talagrand's transportation inequality with an argument similar to Marton's procedure.

Proof of Proposition 3.2: Fix two sets $A, B \subseteq \mathbb{S}^{n-1}$ with $\mu(A), \mu(B) > 0$. Define the cone extension $\bar{A}$ of A as

$$\bar{A} \triangleq \left\{z^n \in \mathbb{R}^n : \frac{z^n}{\|z^n\|} \in A\right\}$$

and define the cone extension $\bar{B}$ of B similarly. It can be easily seen that the measure of A, B under μ are the same as the measures of their cone extensions $\bar{A}, \bar{B}$ under any rotationally invariant probability measure on $\mathbb{R}^n$, and in particular, under the standard Gaussian measure γ, i.e.,

$$\gamma(\bar{A}) = \mu(A) \text{ and } \gamma(\bar{B}) = \mu(B).$$

Now define two conditional probability measures on $\mathbb{R}^n$ based on $\bar{A}, \bar{B}$:

$$\gamma_A(C) \triangleq \frac{\gamma(\bar{A} \cap C)}{\gamma(\bar{A})} \text{ and } \gamma_B(C) \triangleq \frac{\gamma(\bar{B} \cap C)}{\gamma(\bar{B})} \tag{40}$$

for arbitrary $C \subseteq \mathbb{R}^n$. Then $\gamma_A, \gamma_B \ll \gamma$ and we have

$$W_2(\gamma_A, \gamma_B) \le W_2(\gamma_A, \gamma) + W_2(\gamma_B, \gamma) \tag{41}$$

$$\le \sqrt{\mathbb{E}[\|X_A^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2h(X_A^n)}{n}}}} + \sqrt{\mathbb{E}[\|X_B^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2h(X_B^n)}{n}}}} \tag{42}$$

where $X_A^n \sim \gamma_A$ and $X_B^n \sim \gamma_B$, and (41) follows from the triangle inequality and (42) follows from Theorem 2.1. Note that the density function of X_A^n can be expressed as

$$\frac{d\gamma_A}{dx^n}(x^n) = \frac{\mathbf{1}(x^n \in \bar{A})}{\gamma(\bar{A})} \frac{d\gamma}{dx^n}(x^n),$$

and therefore the second moment $\mathbb{E}[\|X_A^n\|^2$ is given by

$$\begin{aligned}
\mathbb{E}_{\gamma_A}[\|X_A^n\|^2] &= \frac{1}{\gamma(\bar{A})} \int_{\mathbb{R}^n} \|x^n\|^2 \mathbf{1}(x^n \in \bar{A}) \gamma(dx^n) \\
&= \frac{1}{\gamma(\bar{A})} \mathbb{E}_\gamma[\|X^n\|^2 \mathbf{1}(X^n \in \bar{A})] \\
&= \frac{1}{\gamma(\bar{A})} \mathbb{E}_\gamma[\mathbf{1}(X^n \in \bar{A})] \mathbb{E}_\gamma[\|X^n\|^2] \qquad (43) \\
&= n \qquad (44)
\end{aligned}$$

and the differential entropy $h(X_A^n)$ is given by

$$\begin{aligned}
h(X_A^n) &= -\mathbb{E}_{\gamma_A}\left[\ln\left(\frac{d\gamma_A}{dx^n}\right)\right] \\
&= -\frac{1}{\gamma(\bar{A})} \int_{\mathbb{R}^n} \ln\left(\frac{1}{\gamma(\bar{A})} \frac{d\gamma}{dx^n}(x^n)\right) \mathbf{1}(x^n \in \bar{A}) \gamma(dx^n) \\
&= \frac{1}{\gamma(\bar{A})} \mathbb{E}_\gamma\left[\left(\frac{n}{2}\ln(2\pi) + \frac{1}{2}\|X^n\|^2 + \ln(\gamma(\bar{A}))\right) \times \mathbf{1}(X^n \in \bar{A})\right] \\
&= \frac{\mathbb{E}_\gamma[\mathbf{1}(X^n \in \bar{A})]}{\gamma(\bar{A})} \mathbb{E}_\gamma\left[\frac{n}{2}\ln(2\pi) + \frac{1}{2}\|X^n\|^2 + \ln(\gamma(\bar{A}))\right] \qquad (45) \\
&= \frac{n}{2}\ln 2\pi e(\gamma(\bar{A}))^{2/n} \\
&= \frac{n}{2}\ln 2\pi e(\mu(A))^{2/n} \qquad (46)
\end{aligned}$$

where both (43) and (45) hold because $\mathbf{1}(X^n \in \bar{A})$ is independent of $\|X^n\|^2$ (except when $X^n = 0$). Similar expressions for $\mathbb{E}[\|X_B^n\|^2]$ and $h(X_B^n)$ can also be obtained and thus (42) simplifies to[3]

$$W_2(\gamma_A, \gamma_B) \le \sqrt{2n(1 - (\mu(A))^{1/n})} + \sqrt{2n(1 - (\mu(B))^{1/n})}. \tag{47}$$

On the other hand, we can also obtain a lower bound on $W_2(\gamma_A, \gamma_B)$. Let $\angle(A, B)$ be the angle distance between A and B, defined as

$$\angle(A, B) \triangleq \inf\{\angle(x^n, y^n) : x^n \in A, y^n \in B\},$$

and assume that $\angle(A, B) \in [0, \pi/2]$ so $\cos(\angle(A, B)) \ge 0$. To lower bound on $W_2(\gamma_A, \gamma_B)$, note that for any coupling P of γ_A and γ_B we have

$$\begin{aligned}
\mathbb{E}_P[\|X_A^n - X_B^n\|^2] &= \mathbb{E}_{\gamma_A}[\|X_A^n\|] + \mathbb{E}_{\gamma_B}[\|X_B^n\|] \\
&\quad - 2\mathbb{E}_P[\|X_A^n\|\|X_B^n\|\cos(\angle(X_A, X_B))] \\
&\ge 2n - 2\mathbb{E}_P[\|X_A^n\|\|X_B^n\|] \cdot \cos(\angle(A, B)) \\
&\ge 2n - 2n\cos(\angle(A, B)) \qquad (48)
\end{aligned}$$

where (48) follows from the Cauchy-Schwarz inequality, and therefore we can get the following lower bound on $W_2(\gamma_A, \gamma_B)$

$$W_2(\gamma_A, \gamma_B) \ge \sqrt{2n - 2n\cos(\angle(A, B))}. \tag{49}$$

Combining this with (47) gives the following inequality:

$$\sqrt{1 - \cos(\angle(A, B))} \le \sqrt{1 - (\mu(A))^{1/n}} + \sqrt{1 - (\mu(B))^{1/n}}. \tag{50}$$

To finish the proof of Proposition 3.2, fix an arbitrary set $A \subseteq \mathbb{S}^{n-1}$ with $\mu(A) = \mu(\text{Cap}(z_0^n, \theta))$ for some arbitrary $z_0^n \in \mathbb{S}^{n-1}$ and $\theta \in (0, \pi/2]$, and choose $B = A_\omega^c = \mathbb{S}^{n-1} \setminus A_\omega$ for $\omega \in (\pi/2 - \theta, \pi/2]$. We will use (50) to show that $\mu(A_\omega^c) \to 0$ as $n \to \infty$. The proof of the proposition for larger ω, follows from the fact that $\mu(A_\omega)$ is increasing in ω. Note that by definition, we have

$$\angle(A, A_\omega^c) = \omega. \tag{51}$$

Plugging this into (50), and also using (38) we obtain

$$\sqrt{1 - \cos\omega} \le \sqrt{1 - \sin\theta} + \liminf_{n\to\infty} \sqrt{1 - (\mu(A_\omega^c))^{1/n}}. \tag{52}$$

Therefore, given $\cos\omega < \sin\theta$, i.e. $\omega > \pi/2 - \theta$, we have

$$\liminf_{n\to\infty} \sqrt{1 - (\mu(A_\omega^c))^{1/n}} > 0. \tag{53}$$

This in turn implies that

$$\mu(A_\omega^c) \to 0 \tag{54}$$

as $n \to \infty$, which completes the proof of Proposition 3.2. ■

[3]Note that applying the original Talagrand's inequality (12) to γ_A and γ_B here would yield $W_2(\gamma_A, \gamma_B) \le \sqrt{2\ln\frac{1}{\mu(A)}} + \sqrt{2\ln\frac{1}{\mu(B)}}$ instead of (47). This inequality is weaker than (47) and follows from (47) by using the fact that $\ln x + 1 \le x$.

C. A New Measure Concentration Result on the Sphere

We next show that the transportation inequality for information constrained OT leads to a new concentration of measure result on $(\mathbb{S}^{n-1}, \mu)$, which recovers Proposition 3.2 as a special case. This new result was recently proved in [13] and [14] by using Riesz' rearrangement inequality [25] and can be stated as follows:

Proposition 3.3: Let $A \subseteq \mathbb{S}^{n-1}$ be an arbitrary set with $\mu(A) = \mu(\text{Cap}(z_0^n, \theta))$ for some arbitrary $z_0^n \in \mathbb{S}^{n-1}$ and $\theta \in (0, \pi/2]$. Then for any $\omega \in (\pi/2 - \theta, \pi/2]$ and $\epsilon > 0$,

$$\mu(\{y^n : \ln \mu(A \cap \text{Cap}(y^n, \omega)) > \ln V(\theta, \omega) - n\epsilon\}) \to 1, \quad (55)$$

in which $V(\theta, \omega)$ is defined as

$$V(\theta, \omega) = \mu(\text{Cap}(z_0^n, \theta) \cap \text{Cap}(y_0^n, \omega)), \quad (56)$$

where z_0^n, y_0^n are perpendicular to each other, i.e. $\angle(z_0^n, y_0^n) = \pi/2$.

In the proposition, $V(\theta, \omega)$ corresponds to the intersection measure of two spherical caps with poles perpendicular to each other. By using the surface area formula for the intersection of two spherical caps in [14, Appendix C-B], one can provide an asymptotic characterization of $\ln V(\theta, \omega)$,

$$\frac{1}{n} \ln V(\theta, \omega) \to \frac{1}{2} \ln(\sin^2 \theta - \cos \omega^2), \text{ as } n \to \infty. \quad (57)$$

Note that an equivalent way to state the blowing-up lemma in Proposition 3.2 is the following: Let $A \subseteq \mathbb{S}^{n-1}$ be an arbitrary set with $\mu(A) = \mu(\text{Cap}(z_0^n, \theta))$ for some arbitrary $z_0^n \in \mathbb{S}^{n-1}$ and $\theta \in (0, \pi/2]$. Then for any $\omega \in (\pi/2 - \theta, \pi/2]$,

$$\mu(\{y^n : A \cap \text{Cap}(y^n, \omega) \neq \emptyset\}) \to 1.$$

This is true because $A \cap \text{Cap}(y^n, \omega) \neq \emptyset$ if and only if $y^n \in A_\omega$. Proposition 3.3 extends Proposition 3.2 by providing a lower bound on $\mu(A \cap \text{Cap}(y^n, \omega))$ for $\omega \in (\pi/2 - \theta, \pi/2]$. When A itself is a cap, (55) is straightforward and follows from the fact that Y^n 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. Proposition 3.3 asserts that this intersection measure is w.h.p. lower bounded by V for any arbitrary A with the same measure. In other words, the spherical cap not only minimizes the measure of its neighborhood as captured by Proposition 3.2, but roughly speaking, also minimizes its intersection measure with the neighborhood of a randomly chosen point on the sphere.

Proof of Proposition 3.3: Fix two sets $A, B \subseteq \mathbb{S}^{n-1}$ with $\mu(A), \mu(B) > 0$. Consider their cone extensions $\bar{A}, \bar{B}$ and the induced conditional probability measures γ_A, γ_B as defined in (40). Since $\gamma_B \ll \gamma$ and γ is absolutely continuous with respect to the Lebesgue measure, the optimal coupling that attains $W_2(\gamma_B, \gamma)$ is a one-to-one mapping that pushes γ forward to γ_B. We will now upper bound $W_2(\gamma_A, \gamma_B; R, \tau, 6n/\tau^2)$ by using a triangle inequality as stated in the following lemma, whose proof is included in Appendix D.

Lemma 3.1: Consider three measures $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\Omega)$ such that there exists a one-to-one mapping $g : \Omega \to \Omega$ satisfying that

1) μ_1 is the push-forward measure of μ_2 under g, i.e., $\mu_1(A) = \mu_2(g^{-1}(A))$ for every Borel set $A \subseteq \Omega$;
2) g induces one optimal coupling that attains $W_p(\mu_1, \mu_2)$, i.e.,

$$W_p(\mu_1, \mu_2) = \{\mathbb{E}_{Y \sim \mu_2}[d^p(g(Y), Y)]\}^{1/p}.$$

Then the following triangle inequality holds:

$$W_p(\mu_1, \mu_3; R, \tau, \delta) \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3; R, \tau, \delta). \quad (58)$$

From the above lemma, for any $R \geq 0$ and $\tau > 0$ we have

$$\begin{aligned} & W_2(\gamma_A, \gamma_B; R, \tau, 6n/\tau^2) \quad (59) \\ & \leq W_2(\gamma_A, \gamma; R, \tau, 6n/\tau^2) + W_2(\gamma_B, \gamma) \\ & \leq \sqrt{\mathbb{E}[\|X_A^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2h(X_A^n)}{n}} \left(1 - e^{-\frac{2R}{n}}\right)}} \\ & \quad + \sqrt{\mathbb{E}[\|X_B^n\|^2] + n - 2n\sqrt{\frac{1}{2\pi e} e^{\frac{2h(X_B^n)}{n}}}} \quad (60) \\ & = \sqrt{2n\left(1 - (\mu(A))^{1/n}\sqrt{1 - e^{-2R/n}}\right)} \\ & \quad + \sqrt{2n(1 - (\mu(B))^{1/n})} \quad (61) \end{aligned}$$

where $X_A^n \sim \gamma_A$ and $X_B^n \sim \gamma_B$; (60) follows from Theorem 2.5 and Theorem 2.1; and (61) follows because $\mathbb{E}[\|X_A^n\|^2] = n$ and $h(X_A^n) = \frac{n}{2} \ln 2\pi e(\mu(A))^{2/n}$ as respectively stated in (44) and (46), and similar expressions hold for $\mathbb{E}[\|X_B^n\|^2]$ and $h(X_B^n)$.

On the other hand, we can also obtain a lower bound on (59). For any $\eta \in [0, \pi]$, let the function $\alpha(\eta)$ be defined as

$$\alpha(\eta) \triangleq \frac{1}{n}\left(\ln \mu(A) \ - R - \sup_{y^n \in B}\{\ln \mu(\text{Cap}(y^n, \eta) \cap A)\}\right) \quad (62)$$

and for any $\epsilon > 0$ define the parameter η_ϵ^* as

$$\eta_\epsilon^* \triangleq \sup\{\eta : \alpha(\eta) \geq \epsilon\}. \quad (63)$$

The following lemma states a lower bound of (59) in terms of η_ϵ^* that will be useful for proving Proposition 3.3. The proof of this lemma will be presented after we finish the proof of Proposition 3.3.

Lemma 3.2: For any $\epsilon > 0$,

$$W_2(\gamma_A, \gamma_B; R, \tau, 6n/\tau^2) \geq \sqrt{2n(1 - \cos \eta_\epsilon^* - \sigma(n, \tau))}$$

where $\sigma(n, \tau) \to 0$ as $\tau/n, n/\tau^2 \to 0$ and $n \to \infty$.

By lemma 3.2 and (61), we get

$$\begin{aligned} \sqrt{1 - \cos \eta_\epsilon^* - \sigma(n, \tau)} \leq & \sqrt{1 - (\mu(A))^{1/n}\sqrt{1 - e^{-2R/n}}} \\ & + \sqrt{1 - (\mu(B))^{1/n}}, \quad (64) \end{aligned}$$

for any $\epsilon > 0$. To finish the proof of Proposition 3.3, fix an arbitrary set $A \subseteq \mathbb{S}^{n-1}$ with $\mu(A) = \mu(\text{Cap}(z_0^n, \theta))$ for some arbitrary $z_0^n \in \mathbb{S}^{n-1}$ and $\theta \in (0, \pi/2]$, and let

$$B \triangleq \{y^n \in \mathbb{S}^{n-1} : \ln \mu(A \cap \text{Cap}\ (y^n, \omega)) \leq \ln V(\theta, \omega) - n\beta\},$$

for some arbitrary $\omega \in (\pi/2-\theta, \pi/2]$ and $\beta > 0$, where $V(\theta,\omega)$ is as defined in (56). In the sequel, we will use (64) to show that $\mu(B) \to 0$ as $n \to \infty$.

To do this, we will apply (64) for a particular choice of $R > 0$ and $\epsilon = \frac{\beta}{4}$. Note that (38) combined with the fact that $\sin\theta > \cos\omega$ implies that

$$\lim_{n\to\infty} (\mu(A))^{1/n} > \cos\omega.$$

This implies that there exists a fixed $\phi > 0$ such that for sufficiently large n

$$(\mu(A))^{1/n} \geq \cos(\omega-\phi). \tag{65}$$

Therefore, letting R be

$$R = \frac{n}{2}\ln\frac{(\mu(A))^{2/n}}{(\mu(A))^{2/n}-\cos^2(\omega-\phi)}, \tag{66}$$

we have that $R > 0$ for n sufficiently large. We will also assume that $\phi > 0$ is chosen sufficiently small so that

$$R \leq \frac{n}{2}\ln\frac{(\mu(A))^{2/n}}{(\mu(A))^{2/n}-\cos^2\omega} + n\frac{\beta}{8}. \tag{67}$$

Note that this is always possible since choosing ϕ smaller makes it easier to satisfy (65). With the choice of R in (66), the first term on the R.H.S. of (64) reduces to

$$\sqrt{1-(\mu(A))^{1/n}\sqrt{1-e^{-2R/n}}} = \sqrt{1-\cos(\omega-\phi)}. \tag{68}$$

Now we will focus on the L.H.S. of (64) and show that it can be lower bounded by

$$\sqrt{1-\cos\eta_\epsilon^* - \sigma(n,\tau)} \geq \sqrt{1-\cos\omega-\sigma(n,\tau)}$$

by choosing $\epsilon = \frac{\beta}{4}$. For this, we evaluate $\alpha(\eta)$ at $\eta = \omega$ under our choice of A, B and R:

$$\begin{aligned}\alpha(\omega) &= \frac{1}{n}\left(-\sup_{y^n\in B}\{\ln\mu(\mathrm{Cap}(y^n,\omega)\cap A)\} + \ln\mu(A) - R\right)\\ &\geq \frac{1}{n}(n\beta - \ln V(\theta,\omega) + \ln\mu(A) - R).\end{aligned} \tag{69}$$

In (69), we can easily lower bound $\ln\mu(A)$ by

$$\ln\mu(A) \geq n\ln\sin\theta - n\frac{\beta}{4} \tag{70}$$

for n sufficiently large. Also, by using (57), we have

$$\ln V(\theta,\omega) \leq \frac{n}{2}\ln(\sin^2\theta-\cos^2\omega) + n\frac{\beta}{4} \tag{71}$$

for n sufficiently large. Moreover, using (38) in (67), R can be further bounded by

$$R \leq \frac{n}{2}\ln\frac{\sin^2\theta}{\sin^2\theta-\cos^2\omega} + n\frac{\beta}{4}, \tag{72}$$

for n sufficiently large. Plugging (70)–(72) into (69), we obtain

$$\alpha(\omega) \geq \frac{\beta}{4},$$

and therefore

$$\omega \leq \eta^*_{\beta/4} \tag{73}$$

by the definition of η_ϵ^* and the nonincreasing property of $\alpha(\eta)$. Hence, by setting $\epsilon = \beta/4$ on the L.H.S. of (64) and using (73), we obtain

$$\sqrt{1-\cos\eta^*_{\beta/4}-\sigma(n,\tau)} \geq \sqrt{1-\cos\omega-\sigma(n,\tau)}. \tag{74}$$

Combining (64), (68) and (74) yields

$$\sqrt{1-\cos\omega-\sigma(n,\tau)} \leq \sqrt{1-\cos(\omega-\phi)} + \sqrt{1-(\mu(B))^{1/n}}.$$

Setting $\tau = n^{3/4}$, we have $\tau/n, n/\tau^2 \to 0$ as $n\to\infty$, and thus $\sigma(n,n^{3/4}) \to 0$ as $n\to\infty$. Therefore, given $\cos\omega < \cos(\omega-\phi)$ and for sufficiently large n, we have

$$\begin{aligned}\mu(B) \leq \Bigg(1 - \Big(&\sqrt{1-\cos\omega-\sigma(n,n^{3/4})}\\ &- \sqrt{1-\cos(\omega-\phi)}\Big)^2\Bigg)^n,\end{aligned} \tag{75}$$

which tends to zero as $n\to\infty$. This completes the proof of Proposition 3.3. ∎

Proof of Lemma 3.2: Consider an arbitrary coupling P of (γ_A,γ_B) that satisfies the $(R,\tau,6n/\tau^2)$-information density constraint. To find a lower bound on $W_2(\gamma_A,\gamma_B;R,\tau,6n/\tau^2)$, it suffices to lower bound $\mathbb{E}_P[\|X_A^n - X_B^n\|^2]$, or equivalently to upper bound $\mathbb{E}_P[X_A^n\cdot X_B^n]$. Fix $\epsilon > 0$ and define

$$F = \{\angle(X_A^n, X_B^n) \geq \eta_\epsilon^*, (X_A^n, X_B^n) \in S\}$$

where

$$\begin{aligned}S \;= \big\{&(x_A^n,x_B^n) : \|x_A^n\|^2 - n| \leq \tau, |\|x_B^n\|^2 - n| \leq \tau,\\ &i_P(x_A^n;x_B^n) \leq R+\tau\big\}.\end{aligned}$$

Then $\mathbb{E}_P[X_A^n\cdot X_B^n]$ can be upper bounded by conditioning on F and F^c respectively, i.e.,

$$\begin{aligned}&\mathbb{E}_P[X_A^n\cdot X_B^n]\\ &= \mathbb{E}_P[X_A^n\cdot X_B^n|F]\mathbb{P}(F) + \mathbb{E}_P[X_A^n\cdot X_B^n|F^c]\mathbb{P}(F^c)\\ &\leq \mathbb{E}_P[X_A^n\cdot X_B^n|F] + \mathbb{E}_P[X_A^n\cdot X_B^n|F^c]\mathbb{P}(F^c).\end{aligned}$$

In the sequel, we will upper bound $\mathbb{E}_P[X_A^n\cdot X_B^n|F]$ and $\mathbb{E}_P[X_A^n\cdot X_B^n|F^c]\mathbb{P}(F^c)$ respectively.

First, from the definition of F, we have

$$\begin{aligned}\mathbb{E}_P[X_A^n\cdot X_B^n|F] &= \mathbb{E}_P[\|X_A^n\|\|X_B^n\|\cos(\angle(X_A^n,X_B^n))|F]\\ &\leq (n+\tau)\cos(\eta_\epsilon^*).\end{aligned} \tag{76}$$

Also, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\mathbb{E}_P[X_A^n\cdot X_B^n|F^c]\mathbb{P}(F^c)\\ &\leq \sqrt{\mathbb{E}_P[\|X_A^n\|^2|F^c]\mathbb{P}(F^c)}\sqrt{\mathbb{E}_P[\|X_B^n\|^2|F^c]\mathbb{P}(F^c)}\\ &= \sqrt{\mathbb{E}[\|X_A^n\|^2] - \mathbb{E}_P[\|X_A^n\|^2|F]\mathbb{P}(F)}\\ &\qquad\times\sqrt{\mathbb{E}[\|X_B^n\|^2] - \mathbb{E}_P[\|X_B^n\|^2|F]\mathbb{P}(F)}\\ &\leq n - (n-\tau)\mathbb{P}(F).\end{aligned} \tag{77}$$

To continue with (77), we need to lower bound $\mathbb{P}(F)$. Since $\mathbb{P}(F)$ can be written as

$$\begin{aligned}\mathbb{P}(F) = \;&\mathbb{P}((X_A^n,X_B^n)\in S)\\ &- \mathbb{P}(\angle(X_A^n,X_B^n) \leq \eta_\epsilon^*, (X_A^n,X_B^n)\in S),\end{aligned}$$

we will bound $\mathbb{P}((X_A^n, X_B^n) \in S)$ and $\mathbb{P}(\angle(X_A^n, X_B^n) \le \eta_\epsilon^*, (X_A^n, X_B^n) \in S)$ respectively.

To bound $\mathbb{P}((X_A^n, X_B^n) \in S)$, note that

$$\begin{aligned}
&\mathbb{P}(|\|X_A^n\|^2 - n| \le \tau) \\
&= \int_{\bar{A}} \mathbf{1}(|\|x_A^n\|^2 - n| \le r)\gamma_A(dx_A^n) \\
&= \frac{1}{\gamma(\bar{A})} \int_{\mathbb{R}^n} \mathbf{1}(|\|x^n\|^2 - n| \le \tau)\mathbf{1}(x^n \in \bar{A})\gamma(dx^n) \\
&= \frac{1}{\gamma(\bar{A})} \mathbb{E}_\gamma[\mathbf{1}(|\|X^n\|^2 - n| \le \tau)\mathbf{1}(X^n \in \bar{A})] \\
&= \frac{1}{\gamma(\bar{A})} \mathbb{E}_\gamma[\mathbf{1}(|\|X^n\|^2 - n| \le \tau)]\mathbb{E}_\gamma[\mathbf{1}(X^n \in \bar{A})] \qquad (78) \\
&= \mathbb{P}(|\|X^n\|^2 - n| \le \tau) \\
&\ge 1 - 3n/\tau^2, \qquad (79)
\end{aligned}$$

where $X^n \sim \gamma$, (78) holds because $\mathbf{1}(|\|X^n\|^2 - n| \le \tau)$ and $\mathbf{1}(X^n \in \bar{A})$ are independent (except when $X^n = 0$), and (79) follows from Chebyshev's inequality. Similarly, $\mathbb{P}(|\|X_B^n\|^2 - n| \le \tau) \ge 1 - 3n/\tau^2$. In addition, since P satisfies the $(R, \tau, 6\tau^2/n)$- information density constraint, we have

$$\mathbb{P}(i_P(X_A^n; X_B^n) \le R + \tau) \ge 1 - 6n/\tau^2.$$

Therefore, by the union bound we have

$$\mathbb{P}((X_A^n, X_B^n) \in S) \ge 1 - 12n/\tau^2. \qquad (80)$$

To upper bound $\mathbb{P}(\angle(X_A^n, X_B^n) \le \eta_\epsilon^*, (X_A^n, X_B^n) \in S)$, we have

$$\begin{aligned}
&\mathbb{P}(\angle(X_A^n, X_B^n) \le \eta_\epsilon^*, (X_A^n, X_B^n) \in S) \\
&= \int_{\bar{B}} \int_{\bar{A}} f_{X_A^n|X_B^n}(x_A^n|x_B^n)\mathbf{1}\left((x_A^n, x_B^n) \in S, \angle(x_A^n, x_B^n) \le \eta_\epsilon^*\right) \\
&\qquad dx_A^n f_{X_B^n}(x_B^n)dx_B^n \\
&\le \int_{\bar{B}} \int_{\bar{A}} e^{R - h(\gamma_A) + \frac{3}{2}\tau}\mathbf{1}((x_A^n, x_B^n) \in S, \angle(x_A^n, x_B^n) \le \eta_\epsilon^*) \\
&\qquad dx_A^n f_{X_B^n}(x_B^n)dx_B^n \qquad (81) \\
&\le \int_{\bar{B}} e^{R - h(\gamma_A) + \frac{3}{2}\tau} e^{-n\epsilon + h(\gamma_A) - R + \frac{1}{2}\tau + n\epsilon_1} f_{X_B^n}(x_B^n)dx_B^n \qquad (82) \\
&= e^{-n(\epsilon - \frac{2\tau}{n} - \epsilon_1)} \\
&\le \epsilon_2, \qquad (83)
\end{aligned}$$

where $\epsilon_1 \to 0$ as $n \to \infty$, and $\epsilon_2 \to 0$ as $n \to \infty$ and $\frac{\tau}{n} \to 0$. In the above, (81) holds because for each $(x_A^n, x_B^n) \in S \cap (\bar{A} \times \bar{B})$, the conditional density $f_{X_A^n|X_B^n}(x_A^n|x_B^n)$ satisfies

$$\begin{aligned}
f_{X_A^n|X_B^n}(x_A^n|x_B^n) &= e^{i_P(x_A^n; x_B^n)} f_{X_A^n}(x_A^n) \\
&\le e^{R+\tau} e^{-\frac{n}{2}\ln(2\pi e\mu(A)^{2/n}) + \frac{1}{2}(n - \|x_A^n\|^2)} \qquad (84) \\
&\le e^{R - h(\gamma_A) + \frac{3}{2}\tau}, \qquad (85)
\end{aligned}$$

where (84) and (85) follows from the facts that $i_P(x_A^n; x_B^n) \le R + \tau$ and $|\|x_A^n\|^2 - n| \le \tau$ respectively by the definition of S. Inequality (82) holds because for each $x_B^n \in \bar{B}$, we have

$$\begin{aligned}
&\int_{\bar{A}} \mathbf{1}((x_A^n, x_B^n) \in S, \angle(x_A^n, x_B^n) \le \eta_\epsilon^n)dx_A^n \\
&\le \mu(A \cap \mathrm{Cap}(x_B^n, \eta_\epsilon^*))|B(0, \sqrt{n+\tau})| \\
&\le \mu(A \cap \mathrm{Cap}(x_B^n, \eta_\epsilon^*))e^{\frac{n}{2}\ln(2\pi e) + \frac{1}{2}\tau + n\epsilon_1} \qquad (86) \\
&\le e^{-n\alpha(\eta_\epsilon^*) + \ln(\mu(A)) - R} e^{\frac{n}{2}\ln(2\pi e) + \frac{1}{2}\tau + n\epsilon_1} \qquad (87) \\
&\le e^{-n\epsilon + \ln(\mu(A)) - R} e^{\frac{n}{2}\ln(2\pi e) + \frac{1}{2}\tau + n\epsilon_1} \qquad (88) \\
&= e^{-n\epsilon + h(\gamma_A) - R + \frac{1}{2}\tau + n\epsilon_1},
\end{aligned}$$

where $|B(0, \sqrt{n+\tau}) = \{x^n : \|x^n\| \le \sqrt{n+\tau}\}|$ denotes the volume of the Euclidean ball with center 0 and radius $\sqrt{n+\tau}$. Here, (86) holds because from [14, Lemma 13], we have

$$|B(0, \sqrt{n+\tau})| \le e^{\frac{n}{2}\ln(2\pi e(1 + \frac{\tau}{n})) + n\epsilon_1} \le e^{\frac{n}{2}\ln 2\pi e + \frac{1}{2}\tau + n\epsilon_1},$$

where the last inequality uses the fact $\ln(1 + a) \le a$ for any $a \ge 0$, (87) follows from the definition of $\alpha(\eta)$, and (88) holds because $\alpha(\eta)$ is continuous in η by Lemma 5.1 and hence $\alpha(\eta_\epsilon^*) \ge \epsilon$.

Combining (80) and (83), we have

$$\begin{aligned}
\mathbb{P}(F) &\ge 1 - 12n/\tau^2 - \epsilon_2 \\
&\ge 1 - \epsilon_3 \qquad (89)
\end{aligned}$$

where $\epsilon_3 \to 0$ as $n \to \infty$, $n/\tau^2 \to 0$ and $\tau/n \to 0$. Combining (76), (77) and (89), we have

$$\mathbb{E}_P[X_A^n \cdot X_B^n] \le n(\cos\eta_\epsilon^* + \sigma(n, \tau)) \qquad (90)$$

where $\sigma(n, \tau) \to 0$ as $n \to \infty$, $n/\tau^2 \to 0$ and $\tau/n \to 0$, and therefore

$$\mathbb{E}_P[\|X_A^n - X_B^n\|^2] \ge 2n(1 - \cos\eta_\epsilon^* - \sigma(n, \tau)).$$

Since the above inequality holds for any coupling P of (γ_A, γ_B) that satisfies the $(R, \tau, 6n/\tau^2)$-information constraint, we can conclude that

$$W_2(\gamma_A, \gamma_B; R, \tau, 6n/\tau^2) \ge 2n(1 - \cos\eta_\epsilon^* - \sigma(n, \tau)).$$

This completes the proof of Lemma 3.2. ■

IV. An Application to Network Information Theory

We next demonstrate an application of our transportation inequalities in network information theory. In particular, we show that the information constrained transportation inequality 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

$$\begin{cases} Z = X + W_1 \\ Y = X + W_2 \end{cases}$$

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 relay 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 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 [14], [16], which shows that $C(C_0)$ can not be 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 [14] and [16] on the capacity of this channel, which builds on a strong data processing inequality (SDPI) for a specific Markov chain. The proof of this SDPI in [14] and [16] is geometric and relies on a packing argument combined with the new measure concentration result stated in Proposition 3.3. We next show that the transportation inequality we develop in the current paper can also be used to establish this SDPI providing a much shorter and simpler proof. In particular, the main idea is to construct an n-letter auxiliary random variable that shares the same marginal distribution with a certain random variable in the relay channel problem but has a different joint distribution (i.e. is coupled differently) with the other random variables in the problem. It is interesting to contrast this converse approach to single-letterization in standard converse programs for network information theory problems. Classically, one uses information measure calculus (e.g., chain rules, non-negativity of divergence) to arrive at single-letter random variables that can be identified as auxiliary random variables. Here, we construct high-dimensional auxiliary random variables whose existence and properties are ensured by the transportation inequality. As we will see in the sequel, this allows us to capture the packing argument employed in [14] and [16] using auxiliary random variables, without the explicit use of geometry or the concentration result in Proposition 3.3. The current method is simpler and may be easier to generalize to other problems.

We now state the above mentioned SDPI and briefly illustrate how it leads to a new upper bound on the relay channel. We then prove it by using the conditional version of the information constrained transportation inequality as stated in Theorem 2.4.

A. A Strong Data Processing Inequality

Consider a long Markov chain

$$Y^n - X^n - Z^n - U_n, \tag{91}$$

with $Z^n = X^n + W_1^n$ and $Y^n = X^n + W_2^n$, where $\mathbb{E}[\|X^n\|^2] = nP$, $W_1^n \sim \mathcal{N}(0, N\ I_n)$, $W_2^n \sim \mathcal{N}(0, I_n)$, and X^n, W_1^n, W_2^n are mutually independent. For this long Markov chain, the following SDPI was established in [14] and [16] and is the key step in resolving Cover's problem.

Proposition 4.1: For the Markov chain described in (91), if $I(Z^n; U_n|Y^n) \leq nC_0$, then $I(X^n; U_n|Y^n)$ is upper bounded by (92), shown at the bottom of the page.

Proposition 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 (91), and

$$\begin{aligned} I(Z^n; U_n|Y^n) &= H(U_n|Y^n) - H(U_n|Z^n, Y^n) \\ &\leq H(U_n) \leq nC_0. \end{aligned}$$

Therefore, by Fano's inequality and Proposition 4.1 we can bound $C(C_0)$ by (93), shown at the bottom of the page, where we have used the simple fact that $I(X^n; Y^n) \leq \frac{n}{2}\ln(1+P)$. The upper bound in (93) resolves Cover's problem as one can easily verify that it is strictly smaller than $n\,C(\infty)$ for any finite C_0.

B. Proof of SDPI via Transportation Inequality

To prove Proposition 4.1, we need the following lemma, which is a consequence of the conditional transportation inequality stated in Theorem 2.4.

Lemma 4.1: For the Markov chain (91), if $I(Z^n; U_n|X^n) = nC'$ for some $C' \geq 0$, 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) $\mathbb{E}[\bar{Z}^n \cdot Y^n] \geq n(P + \sqrt{N(1-e^{-2r})}e^{-C'})$;
3) $I(\bar{Z}^n; Y^n|X^n, U_n) \leq nr$.

Proof: Lemma 4.1 follows immediately from Theorem 2.4 by setting $T = (X^n, U_n)$. In particular, noting that the random vector $W_2^n = Y^n - X^n \sim \mathcal{N}(0, I_n)$ and is independent of (X^n, U_n), we have by Theorem 2.4 that

$$\begin{aligned} &W_2^2(P_{Z^n|X^n,U_n}, P_{Y^n - X^n|X^n,U_n}|P_{X^n,U_n}; nr) \\ &\leq \mathbb{E}[\|Z^n\|^2] + n - 2n\sqrt{1-e^{-2r}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|X^n,U_n)}}. \end{aligned}$$

Therefore, there exists a random vector $\bar{Z}^n$ such that

$$\begin{aligned} &(\bar{Z}^n, X^n, U_n) \sim P_{Z^n, X^n, U_n}, \\ &I(\bar{Z}^n; Y^n|X^n, U_n) \leq nr, \end{aligned}$$

$$I(X^n; U_n|Y^n) \leq \max_{C' \in [0, C_0]} \min_{r>0} \frac{n}{2} \ln \frac{P(N+1-2e^{-C'}\sqrt{N(1-e^{-2r})}) + N(1-e^{-2C'}(1-e^{-2r}))}{(P+1)Ne^{-2r}} \tag{92}$$

$$\begin{aligned} nC(C_0) &\leq I(X^n; Y^n, U_n) + n\epsilon \\ &= I(X^n; Y^n) + I(X^n; U_n|Y^n) + n\epsilon \\ &\leq \max_{C' \in [0, C_0]} \min_{r>0} \frac{n}{2} \ln \frac{P(N+1-2e^{-C'}\sqrt{N(1-e^{-2r})}) + N(1-e^{-2C'}(1-e^{-2r}))}{Ne^{-2r}} + n\epsilon \end{aligned} \tag{93}$$

$$\mathbb{E}[\bar{Z}^n \cdot (Y^n - X^n)] \geq n\sqrt{1-e^{-2r}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|X^n,U_n)}}. \tag{94}$$

This proves 1) and 3) of Lemma 4.1. To show 2) of Lemma 4.1, note that

$$\begin{aligned}
&\mathbb{E}[\bar{Z}^n \cdot Y^n] = \mathbb{E}[\bar{Z}^n \cdot (Y^n - X^n)] + \mathbb{E}[\bar{Z}^n \cdot X^n] \\
&\geq n\sqrt{1-e^{2r}}\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n|X^n,U_n)}} + \mathbb{E}[Z^n \cdot X^n] \qquad (95)\\
&= n\sqrt{1-e^{2r}}\sqrt{Ne^{-2C'}} + nP \qquad (96)\\
&= n(P + \sqrt{N(1-e^{-2r})}e^{-C'})
\end{aligned}$$

where (95) follows from (94) and the fact that $(\bar{Z}^n, X^n) \sim P_{Z^n,X^n}$, and (96) holds because

$$\begin{aligned}
h(Z^n|X^n,U_n) &= h(Z^n|X^n) - I(Z^n;X^n|U_n)\\
&= \frac{n}{2}\ln 2\pi eNe^{-2C'}.
\end{aligned}$$

This completes the proof of Lemma 4.1. ∎

We now use Lemma 4.1 to prove Proposition 4.1. Assuming that for the Markov chain (91),

$$I(Z^n;U_n|X^n) = nC'$$

for some $C' \geq 0$, we can create an auxiliary random vector $\bar{Z}^n$ coupled with X^n, U_n, Y^n so as to satisfy the properties in Lemma 4.1. Therefore, we have

$$\begin{aligned}
&I(X^n;U_n|Y^n)\\
&= I(\bar{Z}^n;U_n|Y^n) + I(X^n;U_n|Y^n,\bar{Z}^n) - I(\bar{Z}^n;U_n|Y^n,X^n)\\
&= I(\bar{Z}^n;U_n|Y^n) + h(U_n|Y^n,\bar{Z}^n) - h(U_n|Y^n,X^n)\\
&\leq I(\bar{Z}^n;U_n|Y^n) + h(U_n|\bar{Z}^n) - h(U_n|X^n)\\
&= I(\bar{Z}^n;U_n|Y^n) - I(\bar{Z}^n;U_n|X^n) \qquad (97)\\
&= h(\bar{Z}^n|Y^n) - h(\bar{Z}^n|Y^n,U_n) - I(\bar{Z}^n;U_n|X^n) \qquad (98)
\end{aligned}$$

where (97) follows because $P_{X^n,\bar{Z}^n,U_n} = P_{X^n,Z^n,U_n}$ by 1) of Lemma 4.1 and thus $X^n - \bar{Z}^n - U_n$ forms a Markov chain. In the following, we will bound the first two terms in (98) respectively. Note that this bounding process precisely mirrors the packing argument used in the geometric proof of [14] and [16], and provides an interpretation of the packing argument in terms of auxiliary random variables.

To bound the first term in (98), we have for any $r > 0$,

$$\begin{aligned}
h(\bar{Z}^n|Y^n) &= h\left(\bar{Z}^n - \frac{\mathbb{E}[\bar{Z}^n \cdot Y^n]}{\mathbb{E}[\|Y^n\|^2]}Y^n \middle| Y^n\right)\\
&\leq h\left(\bar{Z}^n - \frac{\mathbb{E}[\bar{Z}^n \cdot Y^n]}{\mathbb{E}[\|Y^n\|^2]}Y^n\right)\\
&\leq \frac{n}{2}\ln\frac{2\pi e}{n}\mathbb{E}\left[\left\|\bar{Z}^n - \frac{\mathbb{E}[\bar{Z}^n \cdot Y^n]}{\mathbb{E}[\|Y^n\|^2]}Y^n\right\|^2\right]\\
&= \frac{n}{2}\ln\frac{2\pi e}{n}\left(\mathbb{E}[\|\bar{Z}^n\|^2] - \frac{\mathbb{E}[\bar{Z}^n \cdot Y^n]^2}{\mathbb{E}[\|Y^n\|^2]}\right)\\
&\leq \text{R.H.S. of (99)}
\end{aligned}$$

where in the last step we have used 2) of Lemma 4.1. To bound the second term in (98), we have for any $r > 0$,

$$\begin{aligned}
h(\bar{Z}^n|Y^n,U_n) &\geq h(\bar{Z}^n|Y^n,U_n,X^n)\\
&= h(\bar{Z}^n|U_n,X^n) - I(\bar{Z}^n;Y^n|U_n,X^n)\\
&= h(\bar{Z}^n|X^n) - I(\bar{Z}^n;U_n|X^n) - I(\bar{Z}^n;Y^n|U_n,X^n)\\
&\geq \frac{n}{2}\ln 2\pi eN - nC' - nr\\
&= \frac{n}{2}\ln 2\pi Ne^{1-2(C'+r)} \qquad (100)
\end{aligned}$$

where the second inequality follows from 3) of Lemma 4.1.

Plugging (99), shown at the bottom of the page, (100) into (98) gives a bound on $I(X^n;U_n|Y^n)$ in terms of the value of $I(Z^n;U_n|X^n) = nC'$ that holds for any $r > 0$. Therefore, the bound can be tightened by minimizing over $r > 0$. The value C' is unknown, but due to the Markov chain (91) we have

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

The bound in Proposition 4.1 follows by taking a maximum over $C' \in [0, C_0]$.

Appendix A
Comparison of (12) and (13)

Given $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} \ll P_{Y^n}$, let f_{Y^n} and f_{Z^n} denote their respective densities. Then we have

$$\begin{aligned}
&\text{R.H.S. of (12)} = 2\mathbb{E}\left[\ln\frac{f_{Z^n}(Z^n)}{f_{Y^n}(Z^n)}\right]\\
&= 2\mathbb{E}[\ln f_{Z^n}(Z^n)] - 2\mathbb{E}\left[\ln\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{\|Z^n\|^2}{2}\right)\right]\\
&= -2h(Z^n) + n\ln 2\pi + \mathbb{E}[\|Z^n\|^2]\\
&= \mathbb{E}[\|Z^n\|^2] + n - 2\ n\left[\frac{1}{2}\left(\frac{2}{n}h(Z^n) - \ln 2\pi e\right) + 1\right]\\
&= \mathbb{E}[\|Z^n\|^2] + n - 2\ n\left[\ln\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}} + 1\right]\\
&\geq \mathbb{E}[\|Z^n\|^2] + n - 2\ n\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}}\\
&= \text{R.H.S. of (13)}
\end{aligned}$$

where the inequality follows from $\ln a + 1 \leq a$ and holds with equality iff $\sqrt{\frac{1}{2\pi e}e^{\frac{2}{n}h(Z^n)}} = 1$, i.e. $h(Z^n) = \frac{n}{2}\ln 2\pi e$.

Appendix B
On the Tightness of (15)

Here we show that the inequality in (15) is achieved with equality when $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} = \mathcal{N}(\mu, \sigma^2 I_n)$

$$h(\bar{Z}^n|Y^n) \leq \frac{n}{2}\ln 2\pi e\frac{P(N+1-2e^{-C'}\sqrt{N(1-e^{-2r})}) + N(1-e^{-2C'}(1-e^{-2r}))}{P+1} \tag{99}$$

for some μ and σ^2. Specifically, for any coupling $P \in \Pi(P_{Z^n}, P_{Y^n})$ with $I_P(Z^n;Y^n) \le R$, we have

$$\begin{aligned}
R &\ge h(Z^n) - h_P(Z^n|Y^n) \\
&= h(Z^n) - h_P\left(Z^n - \frac{\mathbb{E}_P[Z^n \cdot Y^n]^2}{\mathbb{E}[\|Y^n\|^2]} Y^n \middle| Y^n\right) \\
&\ge h(Z^n) - h_P\left(Z^n - \frac{\mathbb{E}_P[Z^n \cdot Y^n]^2}{\mathbb{E}[\|Y^n\|^2]} Y^n\right) \\
&\ge h(Z^n) - \frac{n}{2}\ln\frac{2\pi e}{n}\mathbb{E}_P\left[\left\|Z^n - \frac{\mathbb{E}_P[Z^n \cdot Y^n]}{\mathbb{E}[\|Y^n\|^2]} Y^n\right\|^2\right] \\
&= h(Z^n) - \frac{n}{2}\ln\frac{2\pi e}{n}\left(\mathbb{E}[\|Z^n\|^2] - \frac{\mathbb{E}_P[Z^n \cdot Y^n]^2}{\mathbb{E}[\|Y^n\|^2]}\right) \\
&= \frac{n}{2}\ln(2\pi e\sigma^2) - \frac{n}{2}\ln 2\pi e\left(\sigma^2 - \frac{\mathbb{E}_P[Z^n \cdot Y^n]^2}{n^2}\right) \\
&= -\frac{n}{2}\ln\left(1 - \frac{\mathbb{E}_P[Z^n \cdot Y^n]^2}{n^2\sigma^2}\right)
\end{aligned}$$

i.e.,

$$\begin{aligned}
\mathbb{E}_P[Z^n \cdot Y^n] &\le n\sqrt{(1 - e^{-\frac{2R}{n}})\sigma^2} \\
&= n\sqrt{\frac{1}{2\pi e} e^{\frac{2}{n}h(Z^n)}(1 - e^{-\frac{2R}{n}})}.
\end{aligned}$$

Therefore, for any coupling $P \in \Pi(P_{Z^n}, P_{Y^n})$ with $I_P(Z^n;Y^n) \le R$, we have

$$\begin{aligned}
\mathbb{E}_P[\|Z^n - Y^n\|^2] &= \mathbb{E}[\|Z^n\|^2] + \mathbb{E}[\|Y^n\|^2] - 2\mathbb{E}_P[Z^n \cdot Y^n] \\
&\ge \text{R.H.S of (15)},
\end{aligned}$$

and thus,

$$\begin{aligned}
W_2^2(P_{Z^n}, P_{Y^n}; R) &= \inf_{\substack{P \in \Pi(P_{Z^n}, P_{Y^n}): \\ I_P(Z^n;Y^n) \le R}} \mathbb{E}_P[\|Z^n - Y^n\|^2] \\
&\ge \text{R.H.S. of (15)}.
\end{aligned}$$

Combining with inequality (15) itself, we can conclude that

$$W_2^2(P_{Z^n}, P_{Y^n}; R) = \text{R.H.S. of (15)}$$

when $P_{Y^n} = \mathcal{N}(0, I_n)$ and $P_{Z^n} = \mathcal{N}(\mu, \sigma^2 I_n)$ for some μ and σ^2.

Appendix C
The Ornstein-Uhlenbeck Semi-Group and Process

The coupling used in the proof of Theorem 2.5 is closely related to the concepts of Ornstein–Uhlenbeck semi-group and Ornstein–Uhlenbeck process. In particular, recall that the Ornstein–Uhlenbeck semi-group is defined as a family of operators $(P_t)_{t\ge 0}$ with

$$P_t f(x_0) = \mathbb{E}[f(X_t)|X_0 = x_0]$$

for all suitable $f : \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}^n$, where $(X_t)_{t\ge 0}$ is a Markov process which admits the explicit representation

$$X_t = e^{-t}(x_0 + \sqrt{2}\int_0^t e^s dB_s) \text{ given } X_0 = x_0,$$

where $(B_t)_{t\ge 0}$ is the standard Brownian motion in $\mathbb{R}^n$ starting at the origin. Such $(X_t)_{t\ge 0}$ is also known as the Ornstein–Uhlenbeck process, and it can be shown to satisfy

$$P_{X_t|X_0=x_0} = e^{-t}x_0 + \sqrt{1 - e^{-2t}}\mathcal{N}(0, I_n), \forall x_0 \in \mathbb{R}^n.$$

Viewed from this perspective, the coupling of (P_{Z^n}, P_{Y^n}) constructed in the proof of Theorem 2.5 can be thought of as the joint distribution of $(g(X_0), X_{t(R)})$ by letting $P_{X_0} = \mathcal{N}(0, I_n)$, g be a mapping that pushes $\mathcal{N}(0, I_n)$ forward to P_{Z^n}, and $t(R) = -\ln\sqrt{1 - e^{-2R/n}}$.

Appendix D
Proof of Lemma 3.1

Let $(X_1, X_2, X_3) \sim P$ be a coupling of (μ_1, μ_2, μ_3) such that

1) $X_1 = g(X_2)$ where g is a one-to-one mapping and $(g(X_2), X_2)$ is an optimal coupling of (μ_1, μ_2) that attains $W_p(\mu_1, \mu_2)$, i.e.,
$$W_p(\mu_1, \mu_2) = \{\mathbb{E}_P[d^p(X_1, X_2)]\}^{1/p};$$
2) (X_2, X_3) is an optimal coupling of (μ_2, μ_3) under the (R, τ, δ)-information density constraint that attains $W_p(\mu_2, \mu_3; R, \tau, \delta)$, i.e.,
$$W_p(\mu_2, \mu_3; R, \tau, \delta) = \{\mathbb{E}_P[d^p(X_2, X_3)]\}^{1/p}.$$

From the above two conditions, it follows that (X_1, X_3) is a coupling of (μ_1, μ_3) that also satisfies the (R, τ, δ)-information density constraint. Indeed, since X_1 and X_2 are one-to-one mappings of each other, we have

$$I_P(X_1; X_3) = I_P(X_2; X_3) \le R$$

and

$$\begin{aligned}
&\mathbb{P}(|i_P(X_1; X_3) - I_P(X_1; X_3)| \le \tau) \\
&= \mathbb{P}(|i_P(X_2; X_3) - I_P(X_2; X_3)| \le \tau) \\
&> 1 - \delta.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&W_p(\mu_1, \mu_3; R, \tau, \delta) \\
&\le \mathbb{E}_P[d(X_1, X_3)^p]^{1/p} \\
&\le \mathbb{E}_P[(d(X_1, X_2) + d(X_2, X_3))^p]^{1/p} \\
&\le \mathbb{E}_P[d(X_1, X_2)^p]^{1/p} + \mathbb{E}_P[d(X_2, X_3)^p]^{1/p} \qquad (101) \\
&= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3; R, \tau, \delta)
\end{aligned}$$

where (101) follows from the Minkowski inequality. This completes the proof of Proposition 3.1.

Appendix E
Continuity of $\alpha(\eta)$

Lemma 5.1: The function $\alpha(\eta)$ defined in (62) is continuous in η.

Proof: Rewrite $\alpha(\eta)$ as

$$\begin{aligned}
&\frac{1}{n}\left(-\sup_{x^n \in B}\{\ln\mu(\text{Cap}(x^n, \eta) \cap A)\} + \ln(\mu(A)) - R\right) \\
&= \frac{1}{n}\left(-\ln\left(\sup_{x^n \in B}\{\mu(\text{Cap}(x^n, \eta) \cap A)\}\right) + \ln(\mu(A)) - R\right).
\end{aligned}$$

To prove $\alpha(\eta)$ is continuous in η, it suffices to show

$$\sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta) \cap A) \tag{102}$$

is continuous in η.

For any $\epsilon > 0$, we have

$$\begin{aligned}
&\left| \sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta+\epsilon) \cap A) - \sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta) \cap A) \right| \\
&= \sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta+\epsilon) \cap A) - \sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta) \cap A) \\
&= \sup_{x^n \in B} \left\{ \mu(\text{Cap}(x^n, \eta+\epsilon) \cap A) - \sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta) \cap A) \right\} \\
&\leq \sup_{x^n \in B} \left\{ \mu(\text{Cap}(x^n, \eta+\epsilon) \cap A) - \mu(\text{Cap}(x^n, \eta) \cap A) \right\} \\
&= \sup_{x^n \in B} \mu((\text{Cap}(x^n, \eta+\epsilon) \setminus \text{Cap}(x^n, \eta)) \cap A) \\
&\leq \sup_{x^n \in B} \mu(\text{Cap}(x^n, \eta+\epsilon) \setminus \text{Cap}(x^n, \eta)) \\
&\leq \delta(\epsilon)
\end{aligned} \tag{103}$$

for some $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and therefore we have shown the right-continuity of (102). Similarly, we can show the left-continuity of (102). This proves the lemma. ■

ACKNOWLEDGMENT

The authors are grateful to Mokshay Madiman and Igal Sason for their helpful discussions and comments. They would also like to thank the anonymous reviewers and the Associate Editor for many valuable comments that helped improve the presentation of this article.

REFERENCES

[1] Y. Bai, X. Wu, and A. Ozgur, "Information constrained optimal transport: From talagrand, to marton, to cover," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2210–2215.

[2] G. Monge, "Memoire sur la theorie des deblais et des remblais," *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

[3] L. V. Kantorovich, "On translation of mass (in Russian), CR," *Doklady. Acad. Sci. (USSR)*, vol. 37, pp. 199–201, Feb. 1942.

[4] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geometric Funct. Anal.*, vol. 6, no. 3, pp. 587–600, May 1996.

[5] Y. Brenier, "Décomposition polaire et réarrangement monotone des champs de vecteurs," *CR Acad. Sci. Paris Ser. I Math.*, vol. 305, pp. 805–808, Jan. 1987.

[6] N. Saldi, T. Linder, and S. Yüksel, "Randomized quantization and source coding with constrained output distribution," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 91–106, Jan. 2015.

[7] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.

[8] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.

[9] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, "Sample complexity of sinkhorn divergences," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1574–1583.

[10] G. Mena and J. Niles-Weed, "Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4541–4551.

[11] K. Marton, "A simple proof of the blowing-up lemma (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 3, pp. 445–446, May 1986.

[12] K. Marton, "Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration," *Ann. Probab.*, vol. 24, no. 2, pp. 857–866, 1996.

[13] L. P. Barnes, A. Ozgur, and X. Wu, "An isoperimetric result on high-dimensional spheres," 2018, *arXiv:1811.10533*.

[14] X. Wu, L. P. Barnes, and A. Ozgur, "The capacity of the relay channel: Solution to cover's problem in the Gaussian case," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 255–275, Jan. 2019.

[15] T. M. Cover, "The capacity of the relay channel," in *Open Problems in Communication and Computation*. Berlin, Germany: Springer, 1987, pp. 72–73.

[16] X. Wu, L. P. Barnes, and A. Ozgur, "The geometry of the relay channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2233–2237.

[17] L. P. Barnes, X. Wu, and A. Ozgur, "A solution to cover's problem for the binary symmetric relay channel: Geometry of sets on the Hamming sphere," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 844–851.

[18] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Found. Trends Commun. Inf. Theory*, vol. 10, nos. 1–2, pp. 1–246, 2013.

[19] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer, 2008.

[20] D. Bakry, F. Bolley, and I. Gentil, "Dimension dependent hypercontractivity for Gaussian kernels," *Probab. Theory Rel. Fields*, vol. 154, nos. 3–4, pp. 845–874, 2012.

[21] O. Rioul and M. H. M. Costa, "On some almost properties," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Jan. 2016, pp. 1–5.

[22] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*. vol. 348. Berlin, Germany: Springer, 2013.

[23] J. Liu and A. Ozgur, "Capacity upper bounds for the relay channel via reverse hypercontractivity," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5448–5455, Sep. 2020.

[24] P. Lévy and F. Pellegrino, *Problémes Concrets D'analyse Fonctionnelle*, vol. 6. Gauthier-Villars Paris, 1951.

[25] A. Baernstein II and B. A. Taylor, "Spherical rearrangements, subharmonic functions, and $*$-functions in N-space," *Duke Math. J.*, vol. 43, no. 2, pp. 245–268, Jun. 1976.

Yikun Bai received the B.Sc. degree in medical imaging from Mudanjiang Medical University, China, in 2012, the M.Sc. degree in mathematics from Marshall University, Huntington, WV, USA, in 2016, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Delaware, in 2021. He is currently a Post-Doctoral Research Assistant at the Machine Intelligence and Neural Technologies (MINT) Laboratory, Vanderbilt University, Nashville, TN, USA. His research interests include machine learning, optimal transport, continual/lifelong learning, and transformation learning.

Xiugang Wu (Member, IEEE) received the B.Eng. degree (Hons.) in electronics and information engineering from Tongji University, Shanghai, China, in 2007, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2009 and 2014, respectively. He was a Post-Doctoral Fellow with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA, from 2015 to 2018. He has been an Assistant Professor at the University of Delaware, Newark, DE, USA, since September 2018, where he is jointly appointed with the Department of Electrical and Computer Engineering and the Department of Computer and Information Sciences. His research interests include information theory, networks, data science, and the interplay between them. He was a recipient of the 2017 NSF Center for Science of Information (CSoI) Postdoctoral Fellowship.

Ayfer Özgür (Member, IEEE) is currently an Associate Professor with the Department of Electrical Engineering, Stanford University, where she is the Chambers Faculty Scholar with the School of Engineering. Her research interests include information theory, wireless communication, statistics, and machine learning. She received the EPFL Best Ph.D. Thesis Award in 2010, the NSF CAREER Award in 2013, the Okawa Foundation Research Grant, the Faculty Research Awards from Google and Facebook, and the IEEE Communication Theory Technical Committee (CTTC) Early Achievement Award in 2018. She was selected as the Inaugural Goldsmith Lecturer of the IEEE ITSoc in 2020.