Katalin Marton's Lasting Legacy

Abbas El Gamal* and Robert M. Gray†

I. Introduction

Katalin Marton (December 9, 1941 − December 13, 2019) was a Hungarian mathematician who made important and lasting contributions to information theory and its applications in probability theory. In 2013 she became the fist woman to win the Claude E. Shannon Award, the IEEE Information Theory Society's highest honor, established in 1972 to recognize profound and consistent contributions to the field. We describe some of her contributions and their lasting impact on the field and on the work of many researchers, including the authors.

The topics covered in this article represent a substantial part of Marton's work but are not comprehensive. We refer the readers to the obituary by Csiszár and Körner [16], her closest collaborators, for additional information about her work and her life.

Marton began her career during a revival in the late 1960s and early 1970s of Shannon's original 1948 and 1959 and Kolmogorov's 1956 work on rate distortion theory. She made several generalizations to Shannon's rate distortion and related results. A decade later and using similar tools, she was a co-founder of the field of distance-divergence inequalities which played a fundamental role in the study of concentration of measure, bringing information theory tools to new audiences. Marton was a pioneer and a major contributor to multiuser information theory, especially the broadcast channel. Her contributions to this problem ranged from defining new classes of broadcast channels to her namesake region, which remains as the tightest known inner bound on the capacity region of this channel. Another notable contribution of Marton to multiuser information theory is her result together with Körner on distributed computation of the modulo 2 sum of two sources. This result was the first to show that structured codes can outperform random codes in multiuser communication and has had significant recent follow-on work.

II. Rate distortion theory

Marton's earliest work, including her first publication [53], dealt with rate distortion theory, a branch of information theory first introduced by Shannon in his classic 1948 paper [83] in which he put forth the general model of a point-to-point communication system depicted in Figure 1. The source X is a random object, which may be a random vector $X^n = (X_1, X_2, \ldots, X_n)$, a discrete time random process $\{X_n : n \in \mathbb{Z}\}$, or a continuous time random process. To reliably communicate the source over the channel, an encoder maps the source outcome into a codeword U, which is then transmitted over the channel. To reproduce the source X, a decoder maps the corresponding channel output V into an estimate $\hat{X}$. Shannon established necessary and sufficient conditions for reliable recovery of the source and showed for several classes of sources and channels the existence of ensembles of codes that can achieve these conditions.

When the channel is noiseless, i.e., $U = V$, with capacity C bits/transmission, the communication model reduces to *source coding* or *compression*. Shannon showed that for a discrete time stationary process drawn from a discrete alphabet, the entropy rate $\overline{H} < C$ is necessary and sufficient for asymptotically lossless reproduction of the source. Motivated by a desire to extend this result to continuous alphabet sources, in Part V of [83] Shannon introduced the notion of *fidelity* in reproducing the source which he quantified as the average "generalized distance" or "cost" — later called a *distortion measure* — between the source vector $x^n \in \mathcal{X}^n$ and its reproduction vector $\hat{x}^n \in \hat{\mathcal{X}}^n$. His primary example was a scalar or *per-letter* distortion measure $d(x_1, \hat{x}_1)$, which implied an additive distortion between the vectors of

$$d_n(x^n, \hat{x}^n) = \sum_{i=1}^{n} d(x_i, \hat{x}_i),$$

with the sum replaced by an integral for continuous time sources. Shannon defined a *fidelity evaluation* of a family of distributions $\{\mathsf{P}_{X^n,\hat{X}^n} : n = 1, 2, \ldots\}$ by the expected distortion $\mathsf{E}_{\mathsf{P}_{X^n,\hat{X}^n}}(d_n(X^n, \hat{X}^n)); n = 1, 2, \ldots$ and the *rate for a source relative to a fidelity criterion*, the rate distortion function of a stationary source, by

$$R_X(D) = \inf_n \frac{1}{n} R_{X^n}(D) = \lim_{n\to\infty} \frac{1}{n} R_{X^n}(D), \qquad (1)$$

$$R_{X^n}(D) = \inf_{\mathsf{P}_{X^n,\hat{X}^n} \in \mathcal{P}_n(D)} I(X^n; \hat{X}^n), \qquad (2)$$

where $I(X^n; \hat{X}^n)$ is the mutual information between X^n and $\hat{X}^n$, and $\mathcal{P}_n(D)$ is the collection of all joint distributions on $(X^n, \hat{X}^n)$ for which the marginal distribution on X^n is P_{X^n}, which can be abbreviated as $\mathsf{P}_{X^n,\hat{X}^n} \Rightarrow \mathsf{P}_{X^n}$, and

$$\frac{1}{n} \mathsf{E}_{\mathsf{P}_{X^n,\hat{X}^n}}(d_n(X^n, \hat{X}^n)) \le D.$$

In Theorem 21 of [83], Shannon showed that when the source is iid, the above expression of $R_X(D)$ reduces to the *single-letter* ($n = 1$) and $R_X(D) < C$ is necessary and sufficient to reproduce the source with the specified distortion, and $R_X(D)$ naturally extends entropy to sources with continuous alphabets. Shannon established the achievability of $R_X(D)$ via random codebook generation using the the distribution resulting from the optimization defining $R_X(D)$ and limiting

* Abbas El Gamal is a professor at Stanford University (email: abbas@ee.stanford.edu). † Robert Gray is an emeritus professor at Stanford University and a research professor at Boston University (email: rmgray@stanford.edu).

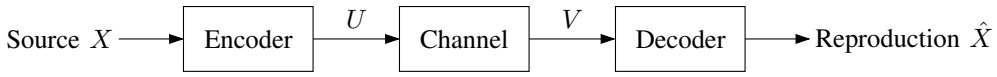

Fig. 1: Shannon's communication system model.

theorems relating sample distortion d_n/n and information density i_n/n to their expectations, where for discrete alphabets

$$i_n(x^n;y^n) = \log \frac{p_{X^n,Y^n}(x^n,y^n)}{p_{X^n}(x^n)p_{Y^n}(y^n)}$$

and in general via the Radon-Nikodym derivative [39][78],

$$i_n(x^n;y^n) = \log \frac{d\,\mathsf{P}_{X^n,Y^n}(x^n,y^n)}{d(\mathsf{P}_{X^n}(x^n) \times \mathsf{P}_{Y^n}(y^n))}.$$

In both cases $\mathsf{E}_{\mathsf{P}_{X^n,Y^n}}(i_n(X^n;Y^n)) = I(X^n;Y^n)$.

When Marton entered the field, Shannon's rate distortion theorem had recently been generalized to stationary and ergodic sources with discrete and abstract alphabets in 1968 by Gallager [23] and in 1971 by Berger [4]. The heart of their general proofs remained the same as Shannon's, but the limiting results were more complicated. The convergence of information density to its expectation generated a new branch of information theory and ergodic theory — commonly called *information stability* or *information ergodic theorems*.

Marton likely learned of Shannon's rate distortion theory [83][84] and contributions by Kolmogorov and his colleagues [39] during her 1969 visit to the Institute of Problems of Information Transmission in Moscow, where Dobrushin and Pinsker both worked. Marton had met the two famous Russian information theorists in Hungary in 1967.

A. Small D asymptotics of $R_X(D)$

In her 1971 paper [53], Marton followed Kolmogorov's [39] "very significant interest ... in the investigations of the asymptotic behavior of the ϵ entropy as $\epsilon \to 0$" and developed upper and lower bounds on the rate distortion function for finite alphabet stationary sources as D approaches zero. Kolmogorov proposed renaming $R_X(D)$ as ϵ-entropy $\overline{H}_\epsilon(X)$, where $\epsilon = D$, to better capture the notion of a generalization of entropy. Her general results are complicated, but in the case of Markov sources they take the simple form

$$\overline{H}(X) - R_X(D) = \frac{1}{m_{r_0}} D \log \frac{1}{D} - O(D),$$

where $m_{r_0} > 0$ is a constant determined by the source and the fidelity criterion.

B. Process definitions of rate distortion functions

Hidden in Lemma 1 of [53] was an elaboration on a suggestion of Kolmogorov [39] — replacing the limits over finite dimensional optimizations of mutual information versus distortion in (1)-(2) by direct optimization over stationary processes In [56] Marton further developed this notion, demonstrating that for stationary sources with complete, separable metric space alphabets and an additive distortion measure, the rate distortion function is

$$R_X(D) = \inf_{\mathsf{P}_{X,\hat{X}} \in \mathcal{P}} \overline{I}(X;\hat{X}),$$

where $\mathcal{P} = \{\mathsf{P}_{X,\hat{X}} : \mathsf{P}_{X,\hat{X}} \Rightarrow \mathsf{P}_X, \mathsf{E}_{\mathsf{P}_{X,\hat{X}}}(d(X_1,\hat{X}_1)) \le D\}$ and

$$\overline{I}(X;\hat{X}) = \lim_{n\to\infty} \frac{1}{n} I(X^n;\hat{X}^n)$$

is the mutual information rate of the stationary process $(X,\hat{X})$. Marton also demonstrated that under some mild additional conditions on the distortion measure, the infimum is a minimum, so there is a stationary pair process that achieves the minimum. Reversing the order of limits and optimization in the definition of the rate distortion function has multiple benefits, primarily adding insight into the meaning of the rate distortion function from a process viewpoint and simplifying previous proofs of Shannon's rate distortion theorem for general sources.

Finding such an optimal stationary pair process provides an alternative approach to the traditional extensions of Shannon's rate distortion theorem beyond iid processes. These traditional proofs require an information stability property for the mutual information densities that yield $R_{X^n}(D)$. In Shannon's original iid case this property followed easily by the classic ergodic theorem. Subsequent generalizations were developed by proving such a property for more general sources. A major difficulty with the general stationary ergodic case was that an ergodic source need not produce ergodic sequences of n-dimensional vectors. Gallager [23] resolved this difficulty using a complicated source decomposition due to Nedoma. Berger [4] used the same technique to further generalize the result to abstract alphabet sources.

Marton's stationary process characterization of the rate distortion function provided an approach to creating a stationary pair process directly with the necessary properties to circumvent the Nedoma decomposition. Others eventually showed that one could also restrict the optimization to ergodic pair processes and that Shannon's approach could result in a generalization of the original iid theorem to stationary and ergodic sources with general alphabets. Marton's result was the catalyst to these simpler and more intuitive results; see, e.g., [34]. Process optimization also played a fundamental role in the development of Orstein's $\overline{d}$ distance [73], which was an important tool in his proof of the isomorphism theorem of ergodic theory establishing the role of equal Shannon entropy rate as necessary and, under additional assumptions, sufficient condition for isomorphism of two random processes. It is also a topic important to Marton's subsequent work.

C. Information stability of stationary and ergodic processes

Marton's 1972 paper [54] extended to stationary ergodic processes a 1963 result of Pinsker [80] that stationary totally ergodic processes possessed the information stability property. This extension provided the key step in generalizing the proof of rate distortion theorem from totally ergodic sources to ergodic sources. Similar results were published in 1971 by

Berger [4] in his generalization of his 1968 paper on the rate distortion theorem for totally ergodic sources. Both Marton and Berger followed Gallager [23] in their use of Nedoma's decomposition.

D. Error exponent for rate distortion for discrete iid sources

In her 1974 paper [55], Marton derived the error exponent for rate distortion coding for discrete iid sources, a result analogous to the behavior of the error probability for discrete memoryless channels studied by Gallager, Shannon, and others. Given an iid source X, an additive distortion, $R > 0, D > 0, \delta > 0$, and a reproduction codebook $\mathcal{C}_n$ of size $\|\mathcal{C}_n\|$, let

$$d_n(x^n, \mathcal{C}_n) = \min_{\hat{x}^n \in \mathcal{C}_n} d_n(x^n, \hat{x}^n)$$

and define the set of "bad" source sequences B_n with respect to $\mathcal{C}_n$ by

$$B_n = \{x^n : \frac{1}{n} d_n(x^n, \mathcal{C}_n) > D + \delta\}.$$

Lastly, define

$$\mathsf{P}^n(R, D) = \min_{\mathcal{C}_n : \|\mathcal{C}_n\| \le 2^{nR}} \mathsf{P}_{X^n}(B_n).$$

Gallager's proof of the source coding theorem shows that for $R > R_{X^n}(D)$, $\lim_{n\to\infty} \mathsf{P}^n(R, D) = 0$. Marton sharpened this result to show that for iid sources the convergence to zero is exponential for a range of $R > R_{X^n}(D)$, and that $\mathsf{P}^n(R, D)$ is otherwise bounded away from zero.

III. Transportation and Measure Concentration

In 1986 Marton [57] introduced a simple inequality comparing a special case of the transportation cost and the informational divergence between two distributions. Optimal transport and the transportation cost date back over two centuries to Monge with its modern revival by Kantorovich. Its rich history and numerous applications in many fields are widely surveyed, see [91] which provides more than 800 references as well as a chapter on concentration of measure that describes Marton's contributions to both topics of this section's title.

Transportation cost has appeared under many names, including Monge (1878), Kantorovich (1942), Ornstein's $\overline{d}$ (d-bar) (1970), Mallows (1972), and Vasershtein/Wasserstein (1969). Informational divergence, also known as relative entropy and Kullback-Leibler divergence, was introduced in 1951 by Kullback and Leibler [46] as a generalization of Shannon's entropy and developed as arguably the most fundamental of the Shannon-style information measures; see, e.g., Pinsker [78] and Kullback [47]. Both notions began with finite-dimensional distributions and were later extended to process distributions.

According to Ledoux [49], the investigation of distance-divergence inequalities, also known as transportation cost-information and transport-entropy inequalities, began in the 1990s with works by Marton [58], [59] and Talagrand [87] in connection with the concentration of measure phenomenon for product measures. The key result by Marton in her 1996 papers in its original finite-dimensional form for discrete sources and Hamming distance appeared a decade earlier [57] in her simple information theoretic proof of the blowing-up lemma, which is discussed later.

A. Transportation and Marton's Inequality

Marton's setup in Lemma 1 of [57] strongly resembles Shannon's. Given a joint distribution $\mathsf{P}_{X^n,\hat{X}^n}$ describing a pair of random vectors $(X^n, \hat{X}^n)$, Shannon optimized mutual information for two cases: fixing a conditional distribution and maximizing over all source distributions P_{X^n} (channel capacity) or fixing the source distribution and minimizing over all conditional distributions satisfying a fidelity criterion $\mathsf{E}_{\mathsf{P}_{X^n,\hat{X}^n}}(d_n(X^n, \hat{X}^n)) \le nD$ (rate distortion). In transportation theory or optimal transportation, the marginal distributions P_{X^n} and $\mathsf{P}_{\hat{X}^n}$ are fixed and the average distortion between them is minimized over the collection $\mathcal{P}_n$ of all joint distributions $\mathsf{P}_{X^n,\hat{X}^n}$ (or couplings) consistent with the given marginals. This defines a transportation cost between finite-dimensional distributions

$$\mathcal{T}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) = \inf_{\mathsf{P}_{X^n,\hat{X}^n} \in \mathcal{P}_n} \mathsf{E}(d_n(X^n, \hat{X}^n)). \quad (3)$$

Sometimes with limits in mind, a normalization $1/n$ is included on the right hand side. The resemblance between the transportation problem and Shannon information optimizations has been exploited in the rate distortion literature, e.g., [33].

The name *optimal transportation* was coined by Kantorovich in 1948 when he realized his cost function introduced in 1942 [37] was equivalent to a 1781 problem of Monge regarding the best way of transporting a pile of dirt of one shape into another. Two special cases are important here: 1) The case of Hamming distance, where (3) becomes Ornstein's $\overline{d}_n$ distance [73], [74]

$$\begin{aligned}\mathcal{T}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) &= \overline{d}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) \\ &= \inf_{\mathsf{P}_{X^n,\hat{X}^n} \in \mathcal{P}_n} \frac{1}{n} \sum_{i=1}^n \mathsf{P}\{X_i \neq \hat{X}_i\}, \end{aligned} \quad (4)$$

and 2) the case of squared-error distortion, where (3) becomes

$$\begin{aligned}\mathcal{T}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) &= W_2^2(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) \\ &= \inf_{\mathsf{P}_{X^n,\hat{X}^n} \in \mathcal{P}_n} \mathsf{E}_{\mathsf{P}_{X^n,\hat{X}^n}} \|X^n - \hat{X}^n\|_2^2. \end{aligned} \quad (5)$$

The square root W_2 of this cost is a metric, commonly known as the Wasserstein distance. It is worth noting here that Wasserstein is the German spelling of the Russian mathematician Leonid Vasershtein, who developed properties of the distance in 1969 which were popularized by Dobrushin [18].

In Lemma 1 of [57], Marton proved the following inequality for random variables with complete separable metric space alphabet and Hamming distance.

Lemma 1. Let $X^n \sim \prod_{i=1}^n \mathsf{P}_{X_i}$ and $\hat{X}^n \sim \mathsf{P}_{\hat{X}^n}$. Then, there exists a joint probability measure $\mathsf{P}_{X^n,\hat{X}^n}$ with these given marginals such that

$$\frac{1}{n} \mathsf{E}(d_n(X^n, \hat{X}^n)) = \frac{1}{n} \sum_{i=1}^n \mathsf{P}\{X_i \neq \hat{X}_i\}$$

$$\leq \left(\frac{1}{n} D\Big(\mathsf{P}_{\hat{X}^n} \Big|\Big| \prod_{i=1}^{n} \mathsf{P}_{X_i} \Big) \right)^{1/2},$$

where $D(\mathsf{P} \,||\, \mathsf{Q}) = \mathsf{E}_{\mathsf{P}} \left(\log(\mathsf{P} \,/\, \mathsf{Q}) \right)$ is the relative entropy between the distributions P and Q.

Hence from (4),

$$\overline{d}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) \leq \left(\frac{1}{n} D\Big(\mathsf{P}_{\hat{X}^n} \Big|\Big| \prod_{i=1}^{n} \mathsf{P}_{X_i} \Big) \right)^{1/2}. \quad (6)$$

Marton later showed in [58], [60] that better constants were possible, but that in general if for some $\rho > 0$ and for all $\mathsf{P}_{\hat{X}^n}$, a distribution P_{X^n} on n-dimensional Euclidean space satisfies the "distance-divergence inequality,"

$$\mathcal{T}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) \leq \left(\frac{2}{\rho} D\Big(\mathsf{P}_{\hat{X}^n} \Big|\Big| \mathsf{P}_{X^n} \Big) \right)^{1/2}, \quad (7)$$

then P_{X^n} has the measure concentration property.

Ornstein [73] developed the process version of the $\overline{d}$ distance as a key tool for proving his isomorphism theorem of ergodic theory: For stationary processes X and $\hat{X}$,

$$\overline{d}(\mathsf{P}_X, \mathsf{P}_{\hat{X}}) = \sup_n \overline{d}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}) = \lim_{n\to\infty} \overline{d}_n(\mathsf{P}_{X^n}, \mathsf{P}_{\hat{X}^n}).$$

A key property of $\overline{d}$ is that it could also be defined directly as the single-letter optimization problem

$$\overline{d}(\mathsf{P}_X, \mathsf{P}_{\hat{X}}) = \overline{d}_1(\mathsf{P}_{X_1}, \mathsf{P}_{\hat{X}_1}). \quad (8)$$

In [58], Marton used the information stability of informational divergence to relate the $\overline{d}$ distance to the relative entropy rate between an arbitrary stationary process $\hat{X}$ and an iid process X,

$$\overline{D}(\mathsf{P}_{\hat{X}} \,\|\, \mathsf{P}_X) \lim_{n\to\infty} \frac{1}{n} D(\mathsf{P}_{\hat{X}^n} \,\|\, \mathsf{P}_{X^n}),$$

to obtain

$$\overline{d}(\mathsf{P}_X, \mathsf{P}_{\hat{X}}) \leq \left(\overline{D}(\mathsf{P}_{\hat{X}} \,\|\, \mathsf{P}_X) \right)^{1/2}. \quad (9)$$

This is the process version of her inequality and was the first inequality to relate limiting transportation cost and relative entropy rate.

Follow-on work. Talagrand [88] extended Marton's inequality in the finite-dimensional case from Hamming distortion to squared-error distortion and Gaussian product measures. This led to his follow-on work on measure concentration from Marton's for Hamming distortion to Gaussian iid processes and squared-error.

Asymptotically optimal source codes resemble fair coin flips. The process version of Marton's inequality (9) easily yields a rigorous proof of a variation on a common intuition of rate distortion theory — if a source coding system has nearly optimal performance, then the bits produced by the encoder should resemble fair coin flips; see, e.g., Problem 10.9 for iid sources in [15]. Gray and Linder [32] established a rigorous result in this direction, with Marton's inequality providing the key step.

Shannon and the bulk of the literature consider block codes. Block codes applied to a stationary ergodic source, however, do not in general produce either a stationary or ergodic channel sequences or reproduction sequence or a combination of the three processes. This complicates drawing conclusions about the properties of the channel process. In his proof of the isomorphism theorem of ergodic theory [74], Ornstein developed techniques for converting block codes into stationary codes (sliding-block or sliding window codes) which inherit useful statistical properties of the block codes they are built upon, such as average distortion and information, and also produce jointly stationary source/encoded source processes which are also ergodic if the source is. Shannon's rate distortion theorem holds for such stationary codes; see, e.g., [31] and the references therein. Pursuing a 1-bit per source sample noiseless channel example, consider the dual Shannon source coding problem of distortion rate theory. The inverse of Shannon's rate distortion function from Marton's process definition [56] is

$$D_X(R) = \inf_{\mathsf{P}_{X,\hat{X}}:\mathsf{P}_{X,\hat{X}} \Rightarrow \mathsf{P}_X, \overline{I}(X;\hat{X}) \leq R} \mathsf{E}(d(X_1, \hat{X}_1)).$$

The stationary codes version of Shannon's source coding theorem implies there exists a sequence of stationary encoder/decoder pairs $f^{(m)}, g^{(m)}$ resulting in a channel process $U^{(m)}$ and reproduction process $\hat{X}^{(m)}$ for which

$$\lim_{m\to\infty} \mathsf{E}_{\mathsf{P}_X}(d(X_1, \hat{X}_1^m)) = D_X(1).$$

If a sequence of encoder-decoder pairs satisfies the Shannon limit, it is said to be *asymptotically optimal*. The original version of Marton's inequality (9) combined with standard information theoretic inequalities easily yields the following.

Theorem [32], [51]. Let X be a stationary ergodic source with Shannon distortion-rate function $D_X(R)$. Let Z denote the process of independent fair coin flips, that is, a Bernoulli process with parameter $1/2$. If sliding-block source codes $f^{(m)}, g^{(m)}$ are asymptotically optimal, then $\lim_{m\to\infty} \bar{d}(\mathsf{P}_{U^{(m)}}, \mathsf{P}_Z) = 0$, that is, the encoder output bits converge to iid coin flips in $\bar{d}$.

Bounds on the capacity of multiuser channels. Marton and Talagrand's distance-divergence inequalities have played a crucial role in the work of Polyanskiy and Wu [81] on the interference channel. They used Talagrand's inequality to resolve the missing corner point problem of Costa for the Gaussian interference channel; see [11], [82], and to obtain an outer bound on its capacity region, and Marton's inequality (6) to establish an outer bound for the discrete memoryless case. Both inequalities provided a critical step in quantifying an approximation error in the output entropy when an input distribution induced by a codebook is replaced by its iid approximation. More recently, Bai, Wu, and Özgür [3], incorporated an information constraint into the optimization defining the transportation cost for the squared-error distortion and generalized Talagrand's distance-divergence inequality for squared-error to this constrained case. This led to further follow-on work to Marton's applications of her inequality, including blowing-up, measure concentration, and Cover's relay channel problem.

B. Blowing-up Lemma

Marton's initial version of her distance-divergence inequality was developed as part of her simple proof for the blowing-up lemma, which was first introduced by Margulis [52] in the study of probabilistic properties of large random graphs. His result was later improved by Ahlswede, Gács, and Körner [1] who used it to establish a strong converse for the degraded broadcast channel by enhancing the known weak converse. Later, Zhang [100] and Xue [98] used it to show that the capacity of the relay channel can be strictly smaller than the cutset bound in [14]. Using simpler and more refined arguments together with the blowing-up lemma, Wu, Özgür, and Xie [97] obtained tighter bounds than in [100], [98].

In [57], Marton provided a significantly simpler information theoretic proof of the blowing-up lemma, which we outline here as a demonstration of her knack for simple and elegant proofs. Let $\mathcal{X}$ be a finite alphabet, $x^n, y^n \in \mathcal{X}^n$, and $d_n(x^n, y^n)$ be the Hamming distance between x^n and y^n. Let $\mathcal{A} \subseteq \mathcal{X}^n$ and for $l \leq n$, let $\Gamma_l(\mathcal{A}) = \{x^n : \min_{y^n \in \mathcal{A}} d_n(x^n, y^n) \leq l\}$ as shown in Figure 2.

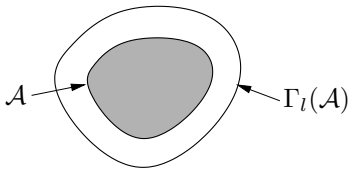

Fig. 2: Two sets in the blowing-up lemma.

Blowing-up Lemma. Let $X^n \sim \mathsf{P}_{X^n} = \prod_{i=1}^n \mathsf{P}_{X_i}$ and $\epsilon_n \to 0$ as $n \to \infty$. There exist $\delta_n, \eta_n \to 0$ as $n \to \infty$ such that if $\mathsf{P}_{X^n}(\mathcal{A}) \geq 2^{-n\epsilon_n}$, then $\mathsf{P}_{X^n}(\Gamma_{n\delta_n}(\mathcal{A})) \geq 1 - \eta_n$.

To prove this lemma, define

$$\mathsf{P}_{\hat{X}^n}(x^n) = \mathsf{P}_{X^n|\mathcal{A}}(x^n) = \begin{cases} \dfrac{\mathsf{P}_{X^n}(x^n)}{\mathsf{P}_{X^n}(\mathcal{A})} & \text{if } x^n \in \mathcal{A}, \\ 0 & \text{if } x^n \notin \mathcal{A}. \end{cases}$$

Then,

$$D\Big(\mathsf{P}_{\hat{X}^n} \Big|\Big| \prod_{i=1}^n \mathsf{P}_{X_i}\Big) = -\log \mathsf{P}_{X^n}(\mathcal{A}) \leq n\epsilon_n.$$

By Lemma 1, there exists $\mathsf{P}_{X^n,\hat{X}^n}$ with the given marginals such that

$$\mathsf{E}(d_n(X^n, \hat{X}^n)) \leq n\sqrt{\epsilon_n}.$$

By the Markov inequality, it follows that

$$\mathsf{P}_{X^n,\hat{X}^n}\{d_n(X^n, \hat{X}^n) \leq n\delta_n\} \geq 1 - \frac{\sqrt{\epsilon_n}}{\delta_n} = 1 - \eta_n,$$

where we choose $\delta_n \to 0$ such that $\eta_n \to 0$ as $n \to \infty$. We therefore have

$$\begin{aligned} \mathsf{P}_{X^n}(\Gamma_{n\delta_n}(\mathcal{A})) &= \mathsf{P}_{X^n,\hat{X}^n}(\Gamma_{n\delta_n}(\mathcal{A}) \times \mathcal{A}) \\ &\quad + \mathsf{P}_{X^n,\hat{X}^n}(\Gamma_{n\delta_n}(\mathcal{A}) \times \mathcal{A}^c) \\ &\overset{(a)}{=} \mathsf{P}_{X^n,\hat{X}^n}(\Gamma_{n\delta_n}(\mathcal{A}) \times \mathcal{A}) \\ &\geq \mathsf{P}_{X^n,\hat{X}^n}\{d_n(X^n, \hat{X}^n) \leq n\delta_n\} \geq 1 - \eta_n, \end{aligned}$$

where (a) follows since $\mathsf{P}_{X^n,\hat{X}^n}(x^n, \hat{x}^n) = 0$ if $\hat{x}^n \notin \mathcal{A}$.

Follow-on work. In a series of papers Paul Shields with Marton developed applications of the blowing-up and related properties to a variety of stationary random processes studied by Ornstein and his colleagues, including weak and very weak Bernoulli processes, and B-processes. See, e.g., [85] and the references therein.

C. Concentration of Measure

Although Marton's initial version of her distance-divergence inequality was developed as part of her simple proof of the blowing-up lemma, which had its original applications in random graph theory and strong converse proofs in multiuser information theory, its major impact came later as a new and powerful tool in the field of concentration of measure. It worth noting, however, that while not originally developed with measure concentration in mind, the blowing-up lemma can be viewed as a concentration inequality.

The theory of concentration of measure arose in the 1970s as a branch of analysis and probability emphasizing the study of asymptotic properties of certain measures and probability distributions, specifically conditions under which a set of positive probability can be expanded very slightly to asymptotically contain most of the probability. At its heart are inequalities that bound, as n grows, the probability of a function of n samples differing from its expectation by more than ϵ as n grows. In Marton's words, "measure concentration is an important property, since it implies sub-Gaussian behavior of the Laplace transforms of Lipschitz functions and thereby is an important tool for proving strong forms of the law of large numbers" [60].

In 1996 Marton [58] used her inequality relating the $\overline{d}$ distance on processes and the relative entropy rate to develop a new approach to proving and extending concentration inequalities, beginning with product distributions (e.g., iid processes) and later extending her results to a class of Markov processes [58], [59], [60].

Follow-on work. Concentration inequalities have proven to be important in analysis, probability, and related areas such as information theory and signal processing. A good survey including history and the contributions of Marton can be found in [49]. As stated earlier, Talagrand [88] extended Marton's inequality to squared-error and used her method to extend the measure conservation property to iid Gaussian processes; see also [60]. In 1997 Dembo [17] used Marton's approach to develop new proofs and sharper versions of many of Talagrand's results on transportation information inequalities. See also Ledoux [49] and Villani [91], who lists several subsequent works following Marton's approach. In his Bibliographic Notes to Chapter 22, Villani [91] states that "It is also Marton who introduced the simple argument by which [transportation] inequalities lead to concentration inequalities…which has since then been reproduced in nearly all introductions to the subject." Boucheron, Lugosi, and Massart [7] also include both a history of concentration inequalities and Marton's role in it along with many applications of her methods by them and others.

IV. The Broadcast channel

In a pioneering paper in multiuser information theory, which aims to extend Shannon's point-to-point information theory to communication networks, Cover [12] introduced the broadcast channel (BC) model depicted in Figure 3. The sender X wishes to reliably communicate a private message $M_i \in [1 : 2^{nR_i}]$ to receiver Y_i, $i = 1, 2$, and a common message $M_0 \in [1 : 2^{nR_0}]$ to both receivers. A rate triple (R_0, R_1, R_2) is said to be achievable if there exists a sequence of codes with this rate triple such that the probability of decoding error can be made as small as desired. The capacity region of this channel is the closure of the set of achievable rate triples; see, e.g., [21] for more details. The capacity region of the broadcast channel is not known in general and is considered one of the most important open problems in the field. Marton jointly with Körner and Csiszár made some of the key early contributions toward solving this problem.

A. Less noisy and more capable

In [12], Cover introduced the technique of *superposition coding* and illustrated it via the binary symmetric BC and Gaussian BC with average transmission power constraint. He then noted that these two example BCs are instances of the class of degraded broadcast channels.

Degraded BC. Given a broadcast channel $p(y_1, y_2|x)$, Y_2 is said to be a degraded version of Y_1 if there exists a channel $p(y_1'|x) = p(y_1|x)$ such that $X \to Y_1' \to Y_2$ form a Markov chain.

Cover conjectured that the capacity region of the degraded BC coincides with the superposition coding region, which we express in the form in [94] that naturally extends to the more general inner bounds discussed later.

Superposition coding inner bound. A rate triple (R_0, R_1, R_2) is achievable for the BC $p(y_1, y_2|x)$ if it satisfies the conditions

$$\begin{aligned} R_2 + R_0 &< I(U; Y_2), \\ R_1 &< I(V; Y_1|U), \\ R_1 + R_2 + R_0 &< I(U, V; Y_1) \end{aligned} \tag{10}$$

for some probability mass function (pmf) $p(u)p(v)$ and function $x(u, v)$, where U and V are auxiliary random variables, that is, random variables that are not defined by the channel itself.

It is not difficult to see that $I(V; Y_1|U) = I(X; Y_1|U)$ and $I(U, V; Y_1) = I(X; Y)$, which yields an equivalent description of the region in (10) for which achievability was settled by Bergmans [5] and the converse was established by Gallager [24] via an ingenious identification of the auxiliary random variable U. Gallager also established a bound on the cardinality of U, which renders the region "computable."

The achievability of the superposition region uses random codebook generation. Fix $p(u)p(v)$ and a function $x(u, v)$. We use $p(u)$ and $p(v)$ to randomly and independently generate codebooks $\{u^n(m_0, m_2) : (m_0, m_2) \in [1 : 2^{nR_0}] \times [1 : 2^{nR_2}]\}$ and $\{v^n(m_1) \; m_1 \in [1 : 2^{n(R_1)}]\}$, respectively. To send (m_0, m_1, m_2), the encoder sends $x(u_i(m_0, m_2), v_i(m_1))$ in transmission $i \in [1 : n]$. Receiver Y_1 uses joint typicality decoding to recover (m_0, m_1, m_2) and receiver Y_2 similarly recovers (m_0, m_2). It can be shown that the decoding error can be made as small as desired if the conditions in (10) are satisfied. One of Marton and Körner's early contributions to the broadcast channel problem [40] was to extend the notion of degradedness to define the following more general classes of BCs for which any message that can be recovered by Y_2 can be (essentially) recovered by Y_1.

Less noisy BC. Given a broadcast channel $p(y_1, y_2|x)$, Y_1 is said to be less noisy than receiver Y_2 if $I(U; Y_1) \geq I(U; Y_2)$ for every $p(u, x)$.

More capable BC. Given a broadcast channel $p(y_1, y_2|x)$, Y_1 is said to be more capable than Y_2 if $I(X; Y_1) \geq I(X; Y_2)$ for every $p(x)$.

It is not difficult to see that the degraded condition implies less noisy, which in turn implies more capable. Körner and Marton showed through examples that these relations are strict. They also showed that superposition coding is optimal for the less noisy BC.

Follow-on work. One of El Gamal's earliest results was to show that superposition coding is optimal for the more capable BC [19]. In [90], van-Dijk introduced the following alternate form of the less noisy condition which proved to be easier to compute and has helped lead to several of the follow-on results discussed later.

Less noisy via concave envelope. Y_1 is less noisy than Y_2 if $(I(X; Y_1) - I(X; Y_2))$ is concave in $p(x)$.

Essentially and effectively less noisy. The definitions of less noisy and more capable BC classes require the mutual information conditions in each case to hold for all distributions. By judiciously restricting the set of distributions over which these conditions apply, extensions of less noisy and more capable for which superposition coding continues to be optimal were introduced.

Motivated by a BC example in which the channel from X to Y_1 is a binary symmetric channel with parameter p and the channel from X to Y_2 is a binary erasure channel with parameter ϵ, Nair [67] introduced the notion of *essentially less noisy* BC. As the parameters of this BC example are varied, it is not difficult to show that: 1) for $0 < \epsilon \leq 2p$, Y_1 is a degraded version of Y_2, 2) for $2p < \epsilon \leq 4p(1-p)$, Y_2 is less noisy than Y_1 but Y_1 is not a degraded version of Y_2, and 3) for $4p(1-p) < \epsilon \leq H(p)$, Y_2 is more capable than Y_1 but not less noisy. For the remaining range, $H(p) < \epsilon < 1$, Nair showed that Y_2 is essentially less noisy than Y_1. Nair also showed that less noisy implies essentially less noisy, but that the reverse does not always hold. He also showed that essentially less noisy does not necessarily imply more capable.

In [38], Kim, Nachman, and El Gamal introduced the notion of *effectively less noisy* and showed that it implies essentially less noisy while being more straightforward to verify computationally. They used this notion along with degraded, less noisy and more capable to show that superposition coding is almost always optimal for the Poisson BC.

More than two receivers. The notions of degraded, less noisy and more capable can be readily extended to BCs with

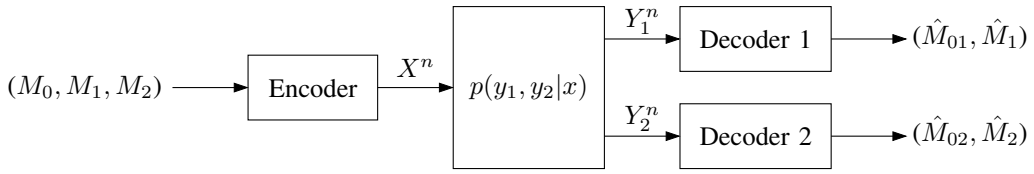

Fig. 3: The two-receiver broadcast channel model.

more than two receivers. It is straightforward to show that superposition coding is optimal for degraded BC with an arbitrary number of receivers. In [64], Nair and Wang showed that superposition coding is also optimal for 3-receiver less noisy BCs. However, in [65] Nair and Xia showed via an example that superposition coding is not necessarily optimal for 3-receiver more capable BCs.

B. Degraded message sets

Consider the broadcast channel in Figure 3 with $M_2 = \emptyset$ or equivalently $R_2 = 0$. In this setting referred to as BC with degraded message sets, the sender wishes to communicate a private message M_1 to receiver Y_1 and a common message M_0 to both receivers. In [41], Körner and Marton showed that the capacity region for this setting coincides with the superposition coding region in (10) with $R_2 = 0$. They established a strong converse using the technique of images of a set [42], which is in itself an important contribution by Marton.

Images of a set. Let $p(y|x)$ be a discrete memoryless channel (DMC) and let $p(x)$ be a pmf over its input alphabet $\mathcal{X}$. For $0 < \epsilon < 1$, let A_n, $n = 1, 2, \ldots$, be a subset of the set of ϵ-typical sequences $\mathcal{T}_\epsilon^{(n)}(p(x))$. The set $B_n \subset \mathcal{Y}^n$ such that $\mathsf{P}(B_n \,|\, x^n) \geq 1 - \epsilon$ for every $x^n \in A_n$ is called the $(1-\epsilon)$-image of A_n under the channel $p(y|x)$. Let $g_n(A_n, 1-\epsilon)$ be the cardinality of the smallest such set. Körner and Marton [42] showed that g_n plays an important role in constructing codes for the DMC.

Maximal code lemma. For sufficiently large n, there exists a codebook for the DMC $p(y|x)$, which is a subset of A_n, with maximal probability of decoding error no greater than ϵ and rate

$$R \geq \frac{1}{n} \log g_n(A_n, 1-\epsilon) - H(Y\,|\,X) - \epsilon.$$

While results of this nature had origins in works as early as 1954 by Feinstein, the main contribution of Körner and Marton was to develop a similar theory for images of a set over two channels with the same input, which involved extensions of the maximal code lemma and new ideas for identification of auxiliary random variables in converse proofs. This enabled them to establish the capacity region of the broadcast channel with degraded message sets and played a central role in their definitions of the classes of less noisy and more capable BCs discussed earlier.

Follow-on work. The notion of degraded message sets can be extended in many ways to more than two receivers. In [6], Borade, Zheng and Trott considered a *multilevel BC* which for the 3-receiver case is defined by $p(y_1, y_3|x)p(y_2|y_1)$, i.e., Y_2 is a degraded version of Y_1, where M_0 is to be communicated to all receivers and M_1 is to be communicated only to receiver Y_1. Nair and El Gamal [68] showed that superposition coding is not in general optimal for this BC and established the capacity region using the technique of indirect coding.

Nair and Wang [69] considered the 3-receiver BC in which M_0 is to be reliably communicated to all receivers and M_1 is to be communicated only to receivers Y_1 and Y_2 and showed that superposition is optimal for several special cases. Nair and Yazdanpanah [66] then showed that superposition coding is not in general optimal for this setting.

C. The Marton region

Marton is most famous in multiuser information theory for establishing the tightest known inner bound on the capacity region of the broadcast channel in [61]. We describe the journey that culminated into the establishment of her inner bound.

In 1975, Cover [13] and van der Meulen [89] independently established the following inner bound on the capacity region of the broadcast channel.

Cover-van der Meulen region. A rate triple (R_0, R_1, R_2) is achievable for the BC $p(y_1, y_2|x)$ if it satisfies the following conditions

$$\begin{aligned} R_0 + R_1 &< I(W, V; Y_1), \\ R_0 + R_2 &< I(W, U; Y_2), \\ R_0 + R_1 + R_2 &< I(W, V; Y_1) + I(U; Y_2|W), \\ R_0 + R_1 + R_2 &< I(V; Y_1|W) + I(W, U; Y_2) \end{aligned} \tag{11}$$

for some pmf $p(w)p(u|w)p(v|w)$ and function $x(w, u, v)$.

It is straightforward to see that by setting $W = U$ and $V = \emptyset$ and with some simple manipulations, we obtain the superposition coding region in (10).

The achievability of the Cover-van der Meulen inner bound involves a new idea beyond the proof for the superposition coding inner bound, which is *rate splitting*. We split M_j, $j = 1, 2$, into two independent parts, a common part $M_{j0} \in [1 : 2^{nR_{j0}}]$ and a private part $M_{jj} \in [1 : 2^{nR_{jj}}]$, hence $R_j = R_{j0} + R_{jj}$. Fix $p(w)p(u|w)p(v|w)$ and a function $x(w, u, v)$ according to $p(w)$. Randomly and independently generate a codebook $\{w^n(m_0, m_{01}, m_{02})\}$, and conditioned on $w^n(m_0, m_{01}, m_{02})$, randomly and conditionally independently generate codebooks $\{u^n(m_0, m_{01}, m_{02}, m_{22})\}$, $\{v^n(m_0, m_{01}, m_{02}, m_{11})\}$ according to $p(u|w)$ and $p(v|w)$, respectively. To send (m_0, m_1, m_2), the encoder sends $x(w_i(m_0, m_{01}, m_{02}), u_i(m_0, m_{o1}, m_{02}, m_{11}),$ $v_i(m_0, m_{o1}, m_{02}, m_{11}))$ at transmission $i \in [1 : n]$. Receiver Y_1 recovers $(m_0, m_{01}, m_{02}, m_{11})$ and receiver Y_2 recovers

$(m_0, m_{01}, m_{02}, m_{22})$. Using standard arguments and Fourier–Motzkin elimination completes the proof; see [21].

The search for a tighter inner bound began with a paper in 1977 by Gelfand [25] in which he established the capacity region of the Blackwell deterministic BC with input $X \in \{0, 1, 2\}$ and outputs $Y_1, Y_2 \in \{0, 1\}$, where for $X = 0$, $Y_1 = Y_2 = 0$, for $X = 1$, $Y_1 = Y_2 = 1$, and $X = 2$, $Y_1 = 0, Y_2 = 1$. In 1978 Pinsker [79] generalized Gelfand's result to establish the capacity region of the deterministic broadcast channel. This result was further generalized to the semideterministic BC independently by Marton [61] and Gelfand and Pinsker [26]. These efforts culminated in the following inner bound on the capacity region of the general BC [61].

Marton region. A rate triple (R_0, R_1, R_2) is achievable for the BC $p(y_1, y_2|x)$ if it satisfies the following conditions

$$\begin{aligned} R_0 + R_1 &< I(W, V; Y_1), \\ R_0 + R_2 &< I(W, U; Y_2), \\ R_0 + R_1 + R_2 &< I(W, V; Y_1) + I(U; Y_2|W) - I(U; V|W), \\ R_0 + R_1 + R_2 &< I(V; Y_1|W) + I(W, U; Y_2) - I(U; V|W), \\ 2R_0 + R_1 + R_2 &< I(W, V; Y_1) + I(W, U; Y_2) - I(U; V|W) \end{aligned} \tag{12}$$

for some pmf $p(w, u, v)$ and function $x(w, u, v)$.

Note that Marton's region can be larger than the Cover—van der Meulen inner bound since the set of distribution on (W, U, V) is larger and the negative term $I(U; V|W)$ drops out of the second and third bounds in (11) if we restrict the set of distributions to the form $p(w)p(u|w)p(v|w)$.

The achievability of the Marton inner bound involves a new key idea beyond the proof of the Cover—van der Meulen bound, which is *multicoding*: Instead of assigning a single codeword to each message, we assign a subcodebook to it from which a codeword is selected for transmission depending on a certain joint typicality condition. This idea was first introduced in 1974 by Kuznetsov and Tsybakov in their paper on computer memory with defects [48]. Multicoding was later used by Gelfand and Pinker [27] and Heegard and El Gamal [36] to generalize the Kuznetsov–Tsybakov result to establish the capacity of the discrete memoryless channel with state known noncausally at the sender. Using this result, Costa in his writing on dirty paper [10] showed that the capacity of the Gaussian channel with additive Gaussian state known noncausally is the same as for the case with no state at all!

Follow-on work. Marton's original proof established achievability for each corner point of the region, then used time sharing to establish achievability for the entire region. Shortly after Marton's result, El Gamal and van der Meulen [22], provided a simple proof that directly establishes achievability for any point in the inner bound without the need for time sharing that we sketch here. Randomly generate a codebook $\{w^n(m_0, m_{01}, m_{02})\}$, and for each $w^n(m_0, m_{01}, m_{02})$, randomly and conditionally independently generate subcodebooks $\{u^n(m_0, m_{01}, m_{02}, l_{22}) : l_{22} \in [1 : r_{22}]\}$ and $\{v^n(m_0, m_{01}, m_{02}, l_{11}) : l_{11} \in [1 : r_{11}]\}$ according to $p(u|w)$ and $p(v|w)$, respectively. To send (m_0, m_1, m_2), find a jointly typical triple $(w^n(m_0, m_{01}, m_{02}), u^n(m_0, m_{01}, m_{02}, l_{22}), v^n(m_0, m_{01}, m_{02}, l_{11}))$ with respect to to $p(w, u, v)$. The conditions for the existence of such a triple are provided by the *mutual covering lemma* [21]. Note that although the messages (M_0, M_1, M_2) are independent of each other, their codewords are generated according to the given $p(w, u, v)$. The details of the proof use standard arguments [22].

El Gamal and Cover [20] subsequently used this mutual covering lemma to establish an inner bound on the rate region of the multiple description coding setting.

MIMO BC. One of the most important follow-on results to Marton's inner bound is proving that it is tight for the multi-input multi-output (MIMO) (or vector Gaussian) BC depicted in Figure 4 with average power transmission constraint $\sum_{i=1}^{n} \mathbf{x}^T(m_0, m_1, m_2, i)\mathbf{x}(m_0, m_1, m_2, i) \le nP$ for $(m_0, m_1, m_2) \in [1 : 2^{nR_0}] \times [1 : 2^{nR_1}] \times [1 : 2^{nR_2}]$. The importance of this result stems from its practical significance as a performance limit on communication over a multiple antenna wireless downlink system.

The first step towards establishing the capacity region of the MIMO BC was by Caire and Shamai [9] who exploited the aforementioned connection between Marton coding and Costa's writing on dirty paper to show that Marton coding is optimal for the sum-capacity of the channel with $r = 2$ and $t = 1$ antennas. The sum-capacity for arbitrary numbers of antennas was independently established by Vishwanath, Jindal, and Goldsmith [92] and Viswanath and Tse [93] using an interesting duality result between a Gaussian BC and a corresponding Gaussian multiple access channel (MAC) with average sum-power constraint, and later by Yu and Cioffi [99] using a minimax argument. The capacity region for the MIMO BC with only private messages, i.e., with $R_0 = 0$, was then established by Weingarten, Steinberg and Shamai [95] using the idea of channel enhancement. An alternative and somewhat simpler proof was given by Mohseni [62]. The capacity region with private and common messages was finally established by Geng and Nair [28] using new powerful techniques for establishing converses for Gaussian channels in general. It is worth pointing out here that the Weingarten–Steinberg–Shamai and the Geng–Nair papers both won the IT Society paper award, a testament to the importance of the Marton region and the exceptional efforts it took to establish these converses.

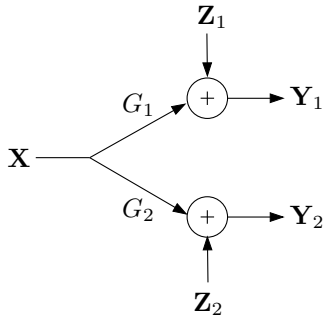

Fig. 4: MIMO BC. The sender $\mathbf{X}$ is a t-dimensional vector, the receivers $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are r-dimensional vectors, G_1, G_2 are $r \times t$ channel gain matrices, and $\mathbf{Z}_1 \sim \mathrm{N}(0, I_r)$ and $\mathbf{Z}_2 \sim \mathrm{N}(0, I_r)$ are the additive noise vectors.

Evaluation of the Marton region. The Marton region as described in (12) is not in general computable since the

auxiliary random variables can have arbitrary cardinalities. The standard method for bounding cardinalities relies on the convexity of the information quantities that are functions of the distributions of the auxiliary random variables; see [21]. The negative term in the description of the Marton region, however, violates this convexity requirement. In [29], Gohari and Ananthram recognized that the cardinality bounds need to hold only at the boundary of the Marton region and devised a *perturbation method* to show that it suffices to consider cardinalities of sizes $|\mathcal{U}|, |\mathcal{V}| \leq |\mathcal{X}|$, $|\mathcal{W}| \leq |\mathcal{X}| + 4$. Using a concave envelope representation of the Marton region, the perturbation method, and a minimax theorem for rate regions, Gohari, Nair and Ananthram [2] obtained the tighter bound $|\mathcal{U}|+|\mathcal{V}| \leq |\mathcal{X}|+1$ for the computation of the optimal weighted sum-rate. This enabled them to show that for binary input BCs, the Marton region reduces to the Cover-van der Meulen region and is achieved using the *randomized time division* strategy developed by Hajek and Pursely in [35].

Optimality of the Marton region. Marton's region is the tightest known inner bound on the capacity of the broadcast channel and is tight for all classes of BCs with known capacity regions. But is it in general tight? There are some indications that it might be. First, efforts to improve upon Marton's inner bound by many researchers over 40 years have failed. These efforts include using simulations to show that the 2-letter version of the Marton region improves upon the 1-letter in (12). Such simulations have been limited, however, because evaluating the region for a BC with input cardinality $|\mathcal{X}| \geq 5$ is computationally intractable. For the very few cases in which simulations were feasible, the 2-letter region did not improve upon the 1-letter. Second, there is no known single outer bound that subsumes all existing outer bounds on the capacity region of the broadcast channel; see, e.g., [30], which suggests that none of them is tight, and perhaps Marton is tight. One of the few general optimality results of the region is by Nair, Kim, and El Gamal [63] who showed that in general, the slope of the region at each corner point is tight.

On the negative side, Padakandla and Pradhan [75] showed using a structured coding scheme that a natural extension of Marton's inner bound to 3-receiver BC is not tight. This suggests that perhaps structured coding, which was pioneered by Marton as discussed in the following section, could improve upon Marton's inner bound even for two receivers.

V. Distributed coding for computing

Consider the distributed source coding for computing setup in Figure 5. Let $(X, Y) \sim p(x, y)$ be two discrete memoryless sources (2-DMS) and $z(x, y)$ be a function of (x, y). The source sequences X^n and Y^n are separately encoded into the indices $M_j \in [1 : 2^{nR_j}]$, $j = 1, 2$, respectively. Upon receiving (M_1, M_2) the decoder finds the estimate $\hat{Z}^n$ of $Z(X_i, Y_i)$, $i = 1, \ldots, n$. The problem is to find the rate region, which is the closure of the set of rate pairs (R_1, R_2) such that the probability of decoding error can be made as small as desired.

For $z(x, y) = (x, y)$, this setting reduces to the distributed source coding problem for which the rate region was established in a pioneering paper on multiuser information theory by Slepian and Wolf [86].

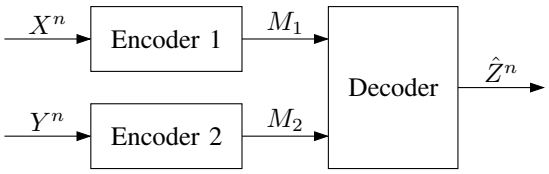

Fig. 5: Distributed source coding for computing setup.

Slepian–Wolf region. The rate region for the distributed source coding of the 2-DMS (X, Y) is the set of rate pairs (R_1, R_2) such that

$$\begin{aligned} R_1 &\geq H(X|Y), \\ R_2 &\geq H(Y|X), \\ R_1 + R_2 &\geq H(X, Y). \end{aligned} \tag{13}$$

This result was quite surprising as it showed that the sum-rate for optimal distributed source coding is the same as for the centralized case in which (X, Y) are encoded together.

In [43], Körner and Marton considered a coding for distributed computing example in which $X \sim \text{Bern}(0.5)$, and $Y = X + Z \mod 2$, where $Z \sim \text{Bern}(p)$, $0 < p < 0.5$, hence the function is $Z = X + Y \mod 2$. The Slepian–Wolf region yields the following inner bound on the rate region for this distributed computing example.

$$\begin{aligned} R_1 &> H(p), \\ R_2 &> H(p), \\ R_1 + R_2 &> 1 + H(p). \end{aligned} \tag{14}$$

It can also be shown that the following is an outer bound on the rate region for this example.

$$R_1 \geq H(p), \quad R_2 \geq H(p). \tag{15}$$

Note that this region can be much larger than (14). Körner and Marton showed using random linear codes that this outer bound is optimal. The idea is to randomly generate a $nR \times n$ binary matrix A and use it to encode X^n and Y^n into the binary nR-sequences $M_1 = AX^n$ and $M_2 = AY^n$, respectively. The decoder then adds these two sequences modulo 2 to obtain AZ^n. By Shannon's asymptotically lossless source coding theorem for a binary iid source via linear codes [21], it can be readily shown that this scheme succeeds provided $R > H(p) + \epsilon$, which yields the region in (15).

Follow-on work. In [72], Orlitsky and Roche provided the following refined outer bound on the rate region of the distributed computing setting

$$R_1 \geq H_{\mathcal{G}_1}(X|Y), \quad R_2 \geq H_{\mathcal{G}_2}(Y|X),$$

where $H_{\mathcal{G}}$ refers to the *graph entropy*, $\mathcal{G}_1$ and $\mathcal{G}_2$ are the characteristic graphs for (X, Y, z) and (Y, X, z), respectively, as defined in [72], and $H_{\mathcal{G}_1}(X|Y) \leq H(X|Y)$ and $H_{\mathcal{G}_2}(Y|X) \leq H(Y|X)$.

The realization that higher rates can be achieved in multiuser communication settings using *structured* rather than "unstructured" random codes led to several follow-on works to [43].

Krithivasan and Pradhan established a rate region for distributed compression of linear functions of correlated Gaussian sources [45] and of discrete memoryless sources [44]. In channel coding settings, the usefulness of structured codes is apparent through the lens of linear network coding in which intermediate nodes forward linear combinations of messages. Nazer and Gastpar showed that noisy multiple access channels could be used for distributed, reliable computation of linear functions first for the modulo-adder case via linear codes [70] and later for the Gaussian case via nested lattice codes [71]. In addition, Wilson et al. [96] utilized structured codes for bidirectional relaying, Bresler, Parekh and Tse [8] used them to approximate the capacity region of the many-to-one Gaussian interference channel to within a constant gap, and Philosof et al. used them to enable distributed dirty paper coding [77].

While the efforts above focused on particular source and channel models for which structured codes are especially well suited, a series of recent papers by Padakandla and Pradhan [76], [75] as well as Lim et al. [50] have shown how multicoding can extend these results to more general settings.

VI. Acknowledgments

The authors would like to thank Michael Gastpar, John Gill, Amin Gohari, Young-Han Kim, Tamás Linder, Chandra Nair and Ayfer Özgür for their valuable contributions to this article.

References

[1] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Probab. Theory Related Fields*, vol. 34, no. 2, pp. 157–177, 1976, correction (1977). 39(4), 353–354.

[2] V. Anantharam, A. Gohari, and C. Nair, "On the evaluation of Marton's inner bound for two-receiver broadcast channels," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1361–1371, 2019.

[3] Y. Bai, , X. Wu, and A. Özgür, "Information constrained optimal transport: From talagrand, to marton, to cover," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020.

[4] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[5] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inf. Theory*, vol. 19, no. 2, pp. 197–207, 1973.

[6] S. Borade, L. Zheng, and M. Trott, "Multilevel broadcast networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Nice, France, 2007, pp. 1151–1155.

[7] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities*. Oxford: Clarendon Press, 2012.

[8] G. Bresler, A. Parekh, and D. Tse, "The approximate capacity of the many-to-one and one-to-many Gaussian interference channels," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4566–4592, 2010.

[9] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[10] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[11] ——, "On the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 31, no. 5, pp. 607–615, 1985.

[12] T. M. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[13] ——, "An achievable rate region for the broadcast channel," *IEEE Trans. Inf. Theory*, vol. 21, no. 4, pp. 399–404, 1975.

[14] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 572–584, 1979.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[16] I. Csiszár and J. Körner, "In memoriam: Katalin Marton 1941–2019," *IEEE Info. Th. Soc. Newsletter*, vol. 70, no. 3, pp. 11–12, 2020.

[17] A. Dembo, "Information inequalities and concentration of measure," *Ann. Probability*, vol. 25, pp. 927–939, 1997.

[18] R. Dobrushin, "Prescribing a system of random variables by conditional distributions," *Theor. Prob. Appl.*, vol. 15, pp. 458–486, 1970.

[19] A. El Gamal, "The capacity of a class of broadcast channels," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 166–169, 1979.

[20] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. 28, no. 6, pp. 851–857, 1982.

[21] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge: Cambridge, 2011.

[22] A. El Gamal and E. C. van der Meulen, "A proof of Marton's coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 120–122, 1981.

[23] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[24] ——, "Capacity and coding for degraded broadcast channels," *Probl. Inf. Transm.*, vol. 10, no. 3, pp. 3–14, 1974.

[25] S. I. Gelfand, "Capacity of one broadcast channel," *Probl. Inf. Transm.*, vol. 13, no. 3, pp. 106–108, 1977.

[26] S. I. Gelfand and M. S. Pinsker, "Capacity of a broadcast channel with one deterministic component," *Probl. Inf. Transm.*, vol. 16, no. 1, pp. 24–34, 1980.

[27] ——, "Coding for channel with random parameters," *Probl. Control Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[28] Y. Geng and C. Nair, "The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2087–2104, 2014.

[29] A. A. Gohari and V. Anantharam, "Evaluation of Marton's inner bound for the general broadcast channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 608–619, 2012.

[30] A. Gohari and C. Nair, "New outer bounds for the two-receiver broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020, pp. 1492–1497.

[31] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 2011, second edition.

[32] R. M. Gray and T. Linder, "Bits in asymptotically optimal lossy source codes are asymptotically Bernoulli," in *Proc. 2009 Data Compression Conference (DCC)*, 2009, pp. 53–62.

[33] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's d-bar distance with applications to information theory," *Ann. Prob.*, vol. 3, pp. 315–328, 1975.

[34] R. M. Gray and F. Saadat, "Block source coding for asymptotically mean stationary sources," *IEEE Trans. Inform. Thy.*, vol. 30, pp. 54–68, 1984.

[35] B. E. Hajek and M. B. Pursley, "Evaluation of an achievable rate region for the broadcast channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 1, pp. 36–46, 1979.

[36] C. Heegard and A. El Gamal, "On the capacity of computer memories with defects," *IEEE Trans. Inf. Theory*, vol. 29, no. 5, pp. 731–739, 1983.

[37] L. V. Kantorovich, "On the translocation of masses," *Dokl. Akad. Nauk*, vol. 37, 1942, English translation in J. Math. Sci. 133, 4 (2006), 1381–1382.

[38] H. Kim, B. Nachman, and A. El Gamal, "Superposition coding is almost always optimal for the Poisson broadcast channel," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1782–1794, 2016.

[39] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inf. Theory*, vol. 2, no. 4, pp. 102–108, 1956.

[40] J. Körner and K. Marton, "Comparison of two noisy channels," in *Topics in Information Theory (Colloquia Mathematica Societatis János Bolyai, Keszthely, Hungary, 1975)*, I. Csiszár and P. Elias, Eds. Amsterdam: North-Holland, 1977, pp. 411–423.

[41] ——, "General broadcast channels with degraded message sets," *IEEE Trans. Inf. Theory*, vol. 23, no. 1, pp. 60–64, 1977.

[42] ——, "Images of a set via two channels and their role in multi-user communication," *IEEE Trans. Inf. Theory*, vol. 23, no. 6, pp. 751–761, 1977.

[43] ——, "How to encode the modulo-two sum of binary sources," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 219–221, 1979.

[44] D. Krithivasan and S. Pradhan, "Distributed source coding using Abelian group codes," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1495–1519, 2011.

[45] D. Krithivasan and S. S. Pradhan, "Lattices for distributed source coding: Jointly Gaussian sources and reconstruction of a linear function," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5268–5651, 2009.

[46] S. Kullback and R. Leibler, "Information and sufficiency," *Ann. of Math. Stat.*, vol. 22, pp. 79–86, 1951.

[47] S. Kullback, *Information theory and statistics*, 2nd ed. Mineola, NY: Dover, 1968.

[48] A. V. Kuznetsov and B. S. Tsybakov, "Coding in a memory with defective cells," *Probl. Inf. Transm.*, vol. 10, no. 2, pp. 52–60, 1974.

[49] M. Ledoux, *The concentration of measure phenomenon*, ser. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001.

[50] S. H. Lim, C. Feng, A. Pastore, B. Nazer, and M. Gastpar, "Compute-forward for DMCs: Simultaneous decoding of multiple combinations," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6242–6255, 2020.

[51] M. Mao, R. M. Gray, and T. Linder, "Rate-constrained simulation and source coding i.i.d. sources," *tit*, vol. 57, no. 7, pp. 4516–4528, 2011.

[52] G. A. Margulis, "Probabilistic characterization of graphs with large connectivity," *Probl. Pered. Inf.*, vol. 10, pp. 101–108, 1974.

[53] K. Marton, "Asymptotics of the epsilon-entropy of discrete stationary processes," *Probl. Peredachi Inf.*, vol. 7, pp. 91–192, 1971.

[54] ——, "Information and information stability of ergodic sources," *Probl. Peredachi Inf.*, vol. 8, pp. 3–8, 1972.

[55] ——, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Thy.*, vol. 20, pp. 197–199, 1974.

[56] ——, "On the rate distortion function of stationary sources," *Problems of Control and Information Theory*, vol. 20, pp. 289–297, 1975.

[57] ——, "A simple proof of the blowing-up lemma (corresp.)," *IEEE Trans. Inf. Theory*, vol. 32, no. 3, pp. 445–446, 1986.

[58] ——, "Bounding $\overline{d}$ distance by informational divergence: a method to prove measure concentration," *Ann. Probab.*, vol. 24, pp. 857–866, 1996.

[59] ——, "A measure concentration inequality for contracting Markov chains," *Geom. Funct. Anal.*, vol. 24, pp. 556–571, 1996.

[60] ——, "Measure concentration for Euclidean distance in the case of dependent random variables," *Annals of Probability*, vol. 32, p. 2526–2544, 2004.

[61] ——, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 3, pp. 306–311, 1979.

[62] M. Mohseni, "Capacity of Gaussian vector broadcast channels," Ph.D. Thesis, Stanford University, Stanford, CA, 2006.

[63] C. Nair, H. Kim, and A. El Gamal, "On the optimality of randomized time division and superposition coding for the broadcast channel," in *Proc. IEEE Inf. Theory Workshop*, 2016, pp. 131–135.

[64] C. Nair and Z. Wang, "The capacity region of the three receiver less noisy broadcast channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4058–4062, 2011.

[65] C. Nair and L. Xia, "On three-receiver more capable channels," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 378–382.

[66] C. Nair and M. Yazdanpanah, "Sub-optimality of superposition coding region for three receiver broadcast channel with two degraded message sets," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 1038–1042.

[67] C. Nair, "Capacity regions of two new classes of two-receiver broadcast channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4207–4214, 2010.

[68] C. Nair and A. El Gamal, "The capacity region of a class of three-receiver broadcast channels with degraded message sets," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4479–4493, 2009.

[69] C. Nair and Z. V. Wang, "On the inner and outer bounds for 2-receiver discrete memoryless broadcast channels," in *Proc. UCSD Inf. Theory Appl. Workshop*, La Jolla, CA, 2008, pp. 226–229. [Online]. Available: http://arxiv.org/abs/0804.3825

[70] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[71] ——, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6463–6486, 2011.

[72] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, 2001.

[73] D. Ornstein, "An application of ergodic theory to probability theory," *Ann. Probab.*, vol. 1, pp. 43–58, 1973.

[74] ——, *Ergodic Theory, Randomness and Dynamical Systems*. New Haven, CT: Yale Univ. Press, 1975.

[75] A. Padakandla and S. S. Pradhan, "Achievable rate region for three user discrete broadcast channel based on coset codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2267–2297, 2018.

[76] ——, "An achievable rate region based on coset codes for multiple access channel with states," *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6393–6415, 2017.

[77] T. Philosof, R. Zamir, U. Erez, and A. J. Khisti, "Lattice strategies for the dirty multiple access channel," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5006–5035, 2011.

[78] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.

[79] ——, "Capacity of noiseless broadcast channels," *Probl. Inf. Transm.*, vol. 14, no. 2, pp. 28–34, 1978.

[80] M. Pinsker, "Sources of messages," *Probl. Peredachi Inf.*, vol. 14, pp. 5–20, 1963.

[81] Y. Poyansky and Y. Wu, "Wasserstein continuity of entropy and outer bounds for interference channels," *IEEE Trans. on Inform. Theor.*, vol. 62, p. 3992, 2016.

[82] I. Sason, "On achievable rate regions for the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1345–1356, 2004.

[83] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 27(4), 623–656, 1948.

[84] ——, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Int. Conv. Rec.*, 1959, vol. 7, part 4, pp. 142–163, reprint with changes (1960). In R. E. Machol (ed.) *Information and Decision Processes*, pp. 93–126. McGraw-Hill, New York.

[85] P. Shields, "The interactions between ergodic theory and information theory," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2079–2093, 1998.

[86] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[87] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Inst. Hautes Études Sci. Publ. Math.*, vol. 81, p. 73–205, 1995.

[88] ——, "Transportation cost for Gaussian and other product measures," *Geometric & Functional Analysis GAFA*, vol. 6, no. 3, pp. 587–600, 1996.

[89] E. C. van der Meulen, "Random coding theorems for the general discrete memoryless broadcast channel," *IEEE Trans. Inf. Theory*, vol. 21, no. 2, pp. 180–190, 1975.

[90] M. van Dijk, "On a special class of broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 712–714, 1997.

[91] C. Villani, *Optimal transport, old and new*. Paris: Springer, 2008.

[92] S. Vishwanath, N. Jindal, and A. J. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.

[93] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[94] L. Wang, E. Sasoglu, B. Bandemer, and Y.-H. Kim, "A comparison of superposition coding schemes," in *Proc. IEEE Int. Symp. Inf. Theory*, 2013, pp. 2970–2974.

[95] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[96] M. P. Wilson, K. Narayanan, H. Pfister, and A. Sprintson, "Joint physical layer coding and network coding for bidirectional relaying," *IEEE Transactions on Information Theory*, vol. 11, no. 56, pp. 5641–5654, 2010.

[97] X. Wu, A. Özgür, and L.-L. Xie, "Improving on the cut-set bound via geometric analysis of typical sets," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2254–2277, 2017.

[98] F. Xue, "A new upper bound on the capacity of a primitive relay channel based on channel simulation," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, p. 4786–4798, 2014.

[99] W. Yu and J. M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, 2004.

[100] Z. Zhang, "Partial converse for a relay channel," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 1106–1110, 1988.