Capacity Approximations for Gaussian Relay Networks

Ritesh Kolte, *Student Member, IEEE*, Ayfer Özgür, *Member, IEEE*, and Abbas El Gamal, *Fellow, IEEE*

***Abstract*— Consider a Gaussian relay network where a source node communicates to a destination node with the help of several layers of relays. Recent work has shown that compress-and-forward-based strategies can achieve the capacity of this network within an additive gap. Here, the relays quantize their received signals at the noise level and map them to random Gaussian codebooks. The resultant gap to capacity is independent of the SNRs of the channels in the network and the topology, but is linear in the total number of nodes. In this paper, we provide an improved lower bound on the rate achieved by the compress-and-forward-based strategies (noisy network coding in particular) in arbitrary Gaussian relay networks, whose gap to capacity depends on the network not only through the total number of nodes but also through the degrees of freedom of the min cut of the network. We illustrate that for many networks, this refined lower bound can lead to a better approximation of the capacity. In particular, we demonstrate that it leads to a logarithmic rather than linear capacity gap in the total number of nodes for certain classes of layered networks. The improvement comes from quantizing the received signals of the relays at a resolution decreasing with the total number of nodes in the network. This suggests that the rule-of-thumb in the literature of quantizing the received signals at the noise level can be highly suboptimal.**

***Index Terms*— Relay networks, gap to capacity, noisy network coding, network topology, quantization.**

I. Introduction

CONSIDER a source node communicating to a destination node via a sequence of relays connected by point-to-point AWGN channels, as depicted in Figure 1. The capacity of this line network is achieved by simple decode-and-forward and is equal to the minimum of the capacities of the successive point-to-point links. The decoding at each stage removes the noise corrupting the information signal and therefore the end-to-end rate achieved is independent of the number of times the message is retransmitted.

Unfortunately, the optimality of decode-and-forward is limited to this line topology, and in physically degraded networks in general. In more general networks with multiple relays at each layer, it is well-understood that the rate achieved by decode-and-forward can be arbitrarily smaller than capacity. Characterizing the capacity of more general networks has been of interest for a long time [3] (also see [4] and references therein). Recently, significant progress has been made ([5]–[9]) which shows that compress-and-forward based strategies can be a better fit for general relay networks. Here, relays quantize/compress their observations without decoding and forward the compressions to the destination by mapping them to a new codebook. In particular, it has been shown that compress-and-forward based relaying strategies (such as quantize-map-and-forward in [5] and noisy network coding in [6]) can achieve rates that are within a bounded gap to the capacity of any relay network with multi-source multicast traffic. The gap is independent of the coefficients and SNR's of the constituent channels and the topology of the network. However, it depends linearly on the total number of nodes which limits the applicability of these results to small networks with a few relays. A recent result that we would like to point out here is [10] in which an extension of partial-decode-forward, called distributed decode-forward, has been shown to achieve a similar result. The gap to capacity for this scheme is also shown to be linear in the number of nodes, with a lower constant compared to noisy network coding.

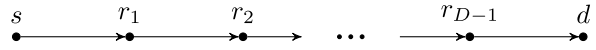

Fig. 1. Line network.

Since the gap to capacity of compress-forward based strategies is linear in the number of nodes, for the line network in Figure 1, they yield an achievable rate whose gap to capacity is linear in the depth of the network D. One natural way to explain this gap is the noise accumulation. As the information signal proceeds deeper into the network, it is corrupted by more and more noise. Therefore, any strategy that does not remove the noise corrupting the signal at each stage by decoding the source message will naturally suffer a rate loss that increases with the number of stages. However, it is not clear why this rate loss should be *linear* in the depth of the network as the current results in the literature suggest [5]–[7]. The total variance of the accumulated noise over the D stages of the network is D times the variance of the noise at each stage (assuming identical noise variances over the D stages). A factor of D increase in the noise variance in a point-to-point Gaussian channel would lead to at most a $\log D$ decrease in capacity, and therefore it is natural to ask if we can reduce the performance loss of compress-and-forward strategies from

Manuscript received July 14, 2014; accepted July 6, 2015. Date of publication July 17, 2015; date of current version August 14, 2015. R. Kolte and A. Özgür were supported in part by the Stanford Graduate Fellowship, in part by the NSF CAREER under Award 1254786, and in part by the NSF Center for Science of Information under Grant CCF-0939370. This paper was presented at the ITW 2013 [1] and IZS 2014 [2].

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: rkolte@stanford.edu; aozgur@stanford.edu; abbas@ee.stanford.edu).

Communicated by T. Liu, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2015.2457904

linear to logarithmic in D, first in the context of this example and then in more general networks.

The first contribution of this paper is to show that a judicious choice of the quantization (or compression) resolutions at the relays can significantly improve the performance of compress-and-forward based strategies (noisy network coding in particular). For example in the line network in Figure 1, if the relay nodes quantize their observed signals at a resolution decreasing linearly in D, the rate loss due to compress-and-forward is only logarithmic in D. (See Section IV.) This is counterintuitive as coarser quantization introduces more noise to the communication and our result suggests that the more relaying stages we have, the more coarsely we should quantize. The rule-of-thumb used in the current literature [5]–[7] is to quantize the received signals at the noise level (independent of the number of relays) which we show to be highly suboptimal. The improvement due to coarser quantization is because in compress-and-forward, there is a rate penalty for communicating the quantized signals to the destination and this rate penalty can be significantly larger than the rate penalty associated with coarser quantization. A detailed discussion on this is presented in Section V. The fact that optimizing the quantization resolutions can lead to better rates for compress-and-forward was also observed in [11] and [12] in the context of the Gaussian diamond network.

An immediate question is whether this observation can lead to better capacity approximations for more general Gaussian networks beyond the line network. To address this question, we suggest a new approximation philosophy for the capacity of Gaussian networks. The current approach is to approximate the capacity within a gap that depends only on the number of nodes. However, two networks with the same number of nodes can have very different topologies which can potentially lead to significantly different performance for compress-and-forward. While it is desirable to have capacity approximations which are independent of the instantaneous channel realizations and SNR's in the network, since these parameters have a wide dynamical range and typically change over a short time scale in wireless networks, topological properties of a network typically change over a much longer time scale. Developing capacity approximations which reveal the dependence of the gap not only on the number of nodes but other structural properties of the network can allow for a better understanding of the performance gap of compress-and-forward strategies as well as yield tighter capacity approximations for many Gaussian networks.

The main result of this paper is a new capacity approximation for Gaussian networks where the gap to capacity depends not only on the number of nodes but also on the number of degrees of freedom (DOF) of the mincut of the network. While the DOF of the mincut of the network can be carefully evaluated for a given network with specific channel realizations (in which case our result will yield the tightest approximation for this network), in many cases this quantity can be easily bounded based only on the topological properties of the network. For example, for the line network in Figure 1 the DOF of the mincut is trivially bounded by 1, while for a diamond network [11] it can be trivially bounded by 2.

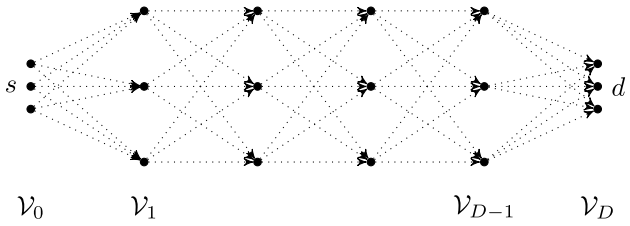

Fig. 2. Multi-layer relay network for $K = 3$, each H_i is a Rayleigh fading matrix.

For such networks, our result yields a logarithmic rather than linear gap in the number of nodes. As before, the improvement is based on a judicious choice of the quantization resolutions at the relays with noisy network coding.

Finally, we look at specific settings and demonstrate that our general result can yield better capacity approximations for these settings than those available in the literature. The first setup we consider is the multi-layer fast-fading Gaussian relay network in Figure 2. Here a source node equipped with K antennas communicates to a destination node equipped with K antennas over D layers, each layer containing K single-antenna relays. Each relay observes a noisy linear combination of the signals transmitted by the relays in the previous layer. All channels are subject to i.i.d. Rayleigh fast-fading. Current results on compress-and-forward [5]–[7] yield a rate which is within $1.3\,KD$ gap to the capacity of this network, where KD is the total number of nodes. Instead, we show that if relays quantize their received signals at a resolution that decreases as the number of layers increases, compress-and-forward can achieve a rate which is within an additive gap of $K \log D + K$ of the network capacity. So for a fixed K, as the number of layers D increases, this gap only grows logarithmically in the depth of the network D.

As a side result, we provide an analysis of the compress-and-forward based strategies in [5]–[7] in fast-fading wireless networks. Fast-fading wireless networks are considered in [5, Th. 8.4], however the conclusion of the theorem and its proof are erroneous. The result in [5, Th. 8.4] suggests that the ergodic fast-fading capacity of a wireless relay network is approximately given by the expected value of the cutset upper bound (where the expectation is over the fading distribution). In contrast, we show that the capacity is approximately given by the minimum of the expected cut values. The difference is in the order of the expectation over the fading distribution and the minimization over different cuts. Note that the second quantity can be arbitrarily larger than the first.

The problem of developing better capacity approximations for this setup has also been considered in [13], where a computation alignment strategy is proposed to remove the accumulating noise with the depth of the network. This yields a gap $7K^3 + 5K \log K$. Computation alignment is based on the idea of combining compute-forward [14] with ergodic alignment proposed in [15]. While the gap to capacity obtained by computation alignment is independent of D, this strategy is significantly more complex than compress-forward and has a number of problems from a practical perspective. In particular,

ergodic alignment over the fading process leads to large delays in communication and requires each relay to know the instantaneous realizations of all the channels in the network. Moreover, its performance critically depends on the symmetry of the fading statistics. The compress-forward strategy with improved quantization we propose in this paper requires only the destination to know the instantaneous channel realizations in the network. In particular, no channel state information is required at the source and at the relays, and the fading statistics are not critical to the operation of the strategy.

To illustrate this last point, we consider another setup where the network has the same layered topology, however the channel coefficients for each link are now fixed with unit magnitudes and arbitrary phases (i.e. each channel coefficient is of the form $e^{j\theta}$ for some arbitrary $\theta \in [0, 2\pi]$). Our approximation gap for this setup is $2K^2 \log D + K \log K + K$ which is again logarithmic in the depth of the network rather than linear. Computation alignment is obviously not applicable in this case and the best currently available capacity approximation for this setup is $1.3KD$ which follows from capacity approximations for general Gaussian networks [5]–[7].

The aforementioned and previous results raise the question of whether tighter gaps scaling sublinearly in the network size can be obtained in the general case (independent of network topology). In this respect, we would like to mention an interesting recent work [16] that shows that obtaining a gap between capacity and cutset bound that is sublinear in the number of nodes for general Gaussian relay networks is possible if and only if the cutset bound is tight for *all* Gaussian relay networks.

The paper is organized as follows. The next section describes the model and some background. The main results and a discussion of the results are presented in Section III. We illustrate the basic idea behind the results via the simple example of a line network in Section IV. Section V aims to clarify the counterintuitive observation that coarser quantization at the relays can result in a better achievable rate. The formal proofs of the main results are presented in Sections VI, VII and VIII.

II. Model and Preliminaries

In the following subsection, we describe the general model of a Gaussian relay network, which is the subject of our main result.

A. General Model

Consider a Gaussian relay network, as depicted in Figure 3 where a source node s communicates to a destination node d a message $m \in [1 : 2^{nR}]$ in n transmissions with the help of a set of relay nodes. Let the number of transmit antennas and receive antennas at node i be M_i and N_i respectively. We assume $N_s = 0$ and $M_d = 0$. Let $\mathcal{N}$ denote the set of all nodes and $M = \sum_{i\in\mathcal{N}} M_i$ and $N = \sum_{i\in\mathcal{N}} N_i$ be the total number of transmit and receive antennas respectively. The signal received by node i at time t is denoted as $\mathbf{Y}_i[t] \in \mathbb{C}^{N_i \times 1}$ which is given by

$$\mathbf{Y}_i[t] = \sum_{j\neq i} \mathbf{H}_{ij}\mathbf{X}_j[t] + \mathbf{Z}_i[t],$$

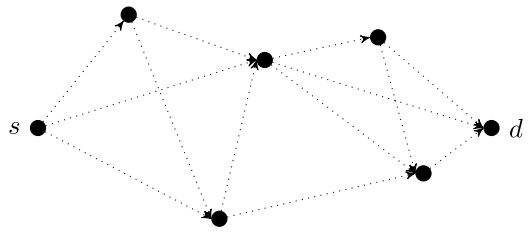

Fig. 3. Gaussian relay network.

where $\mathbf{H}_{ij} \in \mathbb{C}^{N_i \times M_j}$ contains the (complex) channel gains from node j to node i, and $\mathbf{X}_j[t] \in \mathbb{C}^{M_j \times 1}$ is the transmitted vector by node j at time t. We assume that $\mathbf{Y}_s = 0$ and $\mathbf{X}_d = 0$. Each node is subject to an average power constraint P per antenna and $\mathbf{Z}_i[t] \sim \mathcal{CN}(0, \sigma^2 I)$, independent across time and across different receive antennas. The relays are constrained to be strictly causal in their operations, i.e. at any relay node i, $\mathbf{X}_i[t]$ can be a function only of $\{\mathbf{Y}_i[1], \mathbf{Y}_i[2], \ldots, \mathbf{Y}_i[t-1]\}$.

A rate R is said to be achievable if the probability of error of decoding the message $m \in [1 : 2^{nR}]$ at the destination d can be made arbitrarily small by choosing a sufficiently large n. The supremum of all achievable rates is called the capacity C of the network.

In sections VII and VIII, we focus on the following two special cases of Gaussian relay networks respectively.

B. Fast-Fading Layered Network

In section VII, as stated in the introduction and depicted in Figure 2, we consider a fast-fading layered network, where each layer except the first and last contains K single-antenna nodes. The nodes in the ith layer are collectively referred to as $\mathcal{V}_i$ where $0 \leq i \leq D$, while a particular node j in layer i is referred to as the pair (i, j). The layer $\mathcal{V}_0$ consists of the source node s containing K transmit antennas, while the layer $\mathcal{V}_D$ consists of the destination node d, which has K receive antennas. Let $\mathcal{V}^i$ denote $\mathcal{V}_0 \cup \mathcal{V}_1 \cup \cdots \cup \mathcal{V}_i$. We assume that s and d are equipped with multiple antennas in order to keep the problem interesting. Otherwise, the minimum cut becomes the multiple-input-single-output cut from the last layer of relays to d and this trivializes the problem of approximately achieving the capacity of the network. Instead of multiple antennas at d, one can also assume orthogonal bit-pipes from nodes in $\mathcal{V}_{D-1}$ to d, as done in [13].

For $0 \leq i \leq D-1$, the received signal at node $(i+1, j)$ in $\mathcal{V}_{i+1}$ (or antenna if $i = D-1$) depends only on the transmit signals of nodes in $\mathcal{V}_i$ and at time t is given by

$$Y_{(i+1,j)}[t] = \sum_{k=1}^{K} h_{(i,k)\to(i+1,j)}[t] X_{(i,k)}[t] + Z_{(i+1,j)}[t],$$

The channel gain $h_{(i,k)\to(i+1,j)}$ is i.i.d. $\mathcal{CN}(0, 1)$ across time independent of everything else (i.e., other channel gains, noise and transmitted signals). In other words, we assume independent fast Rayleigh fading. The source nodes and the relay nodes do not know the instantaneous realizations of the channel coefficients, i.e have no transmit or receive channel

$$R_{\text{NNC}} \triangleq \sup_{\prod_{k\in\mathcal{N}} p(x_k)p(\hat{y}_k|y_k,x_k)} \min_{\Omega: s\in\Omega, d\in\Omega^c} \Big(I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) - I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c}) \Big). \tag{3}$$

state information. (The source node knows the topology of the network and the channel statistics, i.e. the end-to-end ergodic rate supported by the network.) All instantaneous channel realizations are known at the destination node and are used while decoding the transmitted message from the source node. Thus, we can effectively treat $\{Y_d, H\}$ as the received signal at the destination, where H contains all the channel realizations.

C. Static Layered Network

The topology of the static layered network that we consider in Section VIII is the same as that of the fast-fading layered network, i.e. a source node with K transmit antennas communicates to a destination node with K receive antennas over $D-1$ layers each containing K single-antenna relays. However, instead of assuming fast-fading, we now focus on the case where each channel gain $h_{(i,k)\to(i+1,j)}$ is an arbitrary complex number with unit magnitude, i.e., of the form $e^{j\theta}$ for some arbitrary $\theta \in [0, 2\pi]$ (possibly different for different $(i,k)\to(i+1,j)$), where the j in the superscript stands for the imaginary unit.

D. Background

An upper bound on the capacity C of any relay network is given by the cutset bound [17], which is as follows,

$$C \le \overline{C} \triangleq \sup_{p(x_{\mathcal{N}})} \Big(\min_{\Omega: s\in\Omega, d\in\Omega^c} \overline{C}(\Omega) \Big), \tag{1}$$

where Ω is a subset of $\mathcal{N}$, and

$$\overline{C}(\Omega) \triangleq I(X_\Omega; Y_{\Omega^c}|X_{\Omega^c}), \tag{2}$$

and Ω^c denotes $\mathcal{N}\setminus\Omega$. The notation X_Ω is standard and refers to the set of random variables $\{X_i : i \in \Omega\}$.

In [6], the authors propose an achievability scheme based on compress-and-forward operation at the relays named "noisy network coding" (NNC). This scheme achieves any rate R that is less than R_{NNC}, which is given in (3), as shown at the top of this page. To keep the expressions short, we are assuming that $\hat{Y}_{\Omega^c}$ contains Y_d. In other words, $\hat{Y}_d$ can be set to be equal to Y_d. We refer the reader to [6] for the details of this scheme. It is shown in [6] that the gap between the cutset bound and the rates achieved by noisy network coding for Gaussian relay networks with multi-source multicast traffic is no more than $1.3|\mathcal{N}|$.

III. Main Result

Given a Gaussian relay network as described in Section II-A and a cut of this network $\Omega \subseteq \mathcal{N}$, for any $Q \ge 0$, we define

$$C_Q^{i.i.d.}(\Omega) \triangleq \log\det\left(I + \frac{P}{(Q+1)\sigma^2} \mathbf{H}_{\Omega\to\Omega^c} \mathbf{H}_{\Omega\to\Omega^c}^{\dagger} \right), \tag{4}$$

where the matrix $\mathbf{H}_{\Omega\to\Omega^c}$ denotes the induced MIMO matrix from Ω to Ω^c. In the case of single-antenna nodes, it is obtained by enumerating nodes in Ω and Ω^c in an arbitrary fashion and $\mathbf{H}_{\Omega\to\Omega^c}$ is the $|\Omega^c| \times |\Omega|$ matrix whose (i,j)th entry contains the channel coefficient from node $j \in \Omega$ to node $i \in \Omega^c$. In the case of multiple antennas, it is obtained by enumerating the transmit antennas in Ω and receive antennas in Ω^c and the entries of the matrix denote the corresponding channel coefficient. In this paper, log denotes the natural logarithm. The expression in (4) is the mutual information across the cut Ω, defined in (2), when the channel input distributions at each node are i.i.d. $\mathcal{CN}(0, PI)$ and the noise at each antenna is i.i.d. $\mathcal{CN}(0, (Q+1)\sigma^2)$ (instead of $\mathcal{CN}(0,\sigma^2)$ as originally defined in Section II-A). For a given $Q \ge 0$, let Ω_Q^* be the cut that minimizes $C_Q^{i.i.d.}(\Omega)$,

$$\Omega_Q^* \triangleq \arg\min_{\Omega: s\in\Omega, d\in\Omega^c} C_Q^{i.i.d.}(\Omega). \tag{5}$$

Let d_Q^* be the rank of the corresponding MIMO matrix $\mathbf{H}_{\Omega_Q^*\to(\Omega_Q^*)^c}$. We will also refer to d_Q^* as the number of degrees of freedom of the MIMO channel corresponding to the cut Ω_Q^*, expressed succinctly as

$$d_Q^* = \text{DOF}\left(\arg\min_{\Omega: s\in\Omega, d\in\Omega^c} C_Q^{i.i.d.}(\Omega) \right). \tag{6}$$

Note that the min cut Ω_Q^* and therefore d_Q^* depends on Q. In particular, if Q_1 and Q_2 are two non-negative numbers and say $Q_1 > Q_2 \ge 0$, then $d_{Q_1}^*$ can be larger than, smaller than or same as $d_{Q_2}^*$. The following theorem states our main result.

Theorem 1: The capacity C of the network described in Section II-A satisfies

$$\overline{C} \ge C \ge \overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - \frac{N}{Q} - d_Q^* \log(Q+1),$$

for any non-negative Q, where $\overline{C}$ is the cutset bound of the network given in (1).

Note that Q in the theorem is a free parameter that can be optimized for a given network to minimize the gap between the achieved rate and the cutset upper bound. In the proof of the theorem, we will see that Q corresponds to the variance of the quantization noise introduced at the relays in noisy network coding [6]; larger Q corresponds to coarser quantization. In previous works [5], [6], Q is chosen to be constant independent of the number of nodes (or antennas) N (i.e. $Q \approx 1$ and the quantization noise $Q\sigma^2$ is of the order of the Gaussian noise variance σ^2). Observe that due to the third term N/Q of the gap in Theorem 1, this results in a gap that is at least linear in N. Trivially upper bounding both d_0^* and d_Q^* by N makes the first and the third term also linear in N. However, in many cases, the min cut of the network can have much smaller DOF than M and N and in such cases allowing Q to depend on N can result in a much smaller gap.

For example, in the diamond network with single-antenna at each node it is clear a priori that any cut of the network

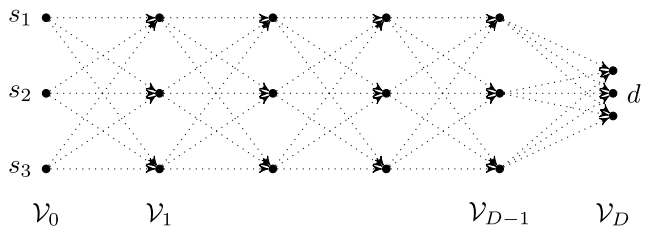

Fig. 4. Fast-fading layered network with multiple sources.

has at most two degrees of freedom, regardless of the number of relays, and therefore $d^*_Q \leq 2$ for any Q. It can be seen immediately from the above theorem that choosing $Q = N$ in this case results in a gap logarithmic in N [11], which compares favorably with a gap that is linear in N. Similarly, for the fast-fading layered network with K single-antenna nodes per layer defined in Section II-B, we show in Section VII that $d^*_Q \leq K$ for any Q. If there are D layers in the network so that $N = M = KD$, the above expression tells us that choosing Q to be proportional to D gives a gap that is logarithmic in D instead of linear in D. In Section VIII, we demonstrate yet another setting in which applying Theorem 1 and choosing Q to be proportional to the number of layers allows us to obtain an improved gap. This demonstrates that the rule of thumb in the current literature to quantize received signals at the noise level ($Q \approx 1$) can be highly suboptimal.

Theorems 2 and 3 stated below provide formally the results that are mentioned in the preceding paragraph.

Theorem 2: The capacity C of the fast-fading layered network described in Section II-B satisfies

$$\overline{C} \geq C \geq \overline{C} - K \log D - K. \tag{7}$$

Theorem 2 follows from evaluating the required quantities in the expression in Theorem 1 for the setup in Section II-B. However, directly applying the result of Theorem 1 for this setup yields a gap of $2K \log D + K$. It turns out that we can further tighten the gap to $K \log D + K$ based on the observation that for this setup, the cutset bound can be evaluated explicitly and the optimal channel input distribution turns out to be independent across the antennas. The detailed proof appears in Section VII-A and VII-B.

The following corollary extends the result of Theorem 2 to the setup considered in [13]. In this setup, instead of a single K-antenna source, there are K single-antenna sources $\{s_1, s_2 \ldots, s_K\}$ interested in communicating with the destination, as depicted in Figure 4. We show that Theorem 2 also implies a similar result for the sum-capacity C of this network.

Corollary 1: The sum-capacity C of the network in Figure 4 satisfies

$$\overline{C} \geq C \geq \overline{C} - K \log D - K. \tag{8}$$

The proof of Corollary 1 appears in Section VII-C.

The following theorem states the result for the static layered network setup, and the proof is given in Section VIII.

Theorem 3: For $K \geq 2$ and $D \geq 2$, the capacity C of the layered network described in Section II-C satisfies

$$\overline{C} \geq C \geq \overline{C} - 2K^2 \log D - K \log K - K. \tag{9}$$

IV. Line Network

We first illustrate the main idea of this paper in a simple setting, the line network in Figure 1. Here we assume that each link i is a AWGN channel with gain h_i and the channel gains h_i are fixed and known. Each node has power P and the noise variance is σ^2. (The conclusions below also hold under a fast-fading assumption similar to the one described in Section II.) It is clear that a decode-forward strategy at the relays achieves the capacity of this line network, while compress-and-forward based strategies (such as quantize-map-forward in [5] and noisy network coding in [6]) with quantization done at the noise level have a gap to capacity that is linear in the number of nodes D. Here, we show that if relays instead quantize at resolution $(D-1)$ times the noise level, the gap to capacity becomes logarithmic in D.

Number the nodes s through d as $0, 1, 2, \ldots, D$. Let's consider the rate achievable by noisy network coding for this network, assuming all relay nodes choose their transmission codebooks independently from a Gaussian distribution, i.e. $X_i \sim \mathcal{CN}(0, P)$ and independent of each other. As described in Section II-D, the rate

$$\min_{0 \leq i \leq D-1} \Big(I(X_i; \hat{Y}_{i+1} | X_{i+1}) - I(Y_{\mathcal{V}^i}; \hat{Y}_{\mathcal{V}^i} | X_{\mathcal{N}}, \hat{Y}_{\mathcal{N} \setminus \mathcal{V}^i}) \Big),$$

is achievable, where $\mathcal{V}^i = \{0, \ldots, i\}$, and each relay chooses $\hat{Y}_i = Y_i + \hat{Z}_i$ where $\hat{Z}_i \sim \mathcal{N}(0, (D-1)\sigma^2)$ independent of everything else. Since $Y_{i+1} = h_i X_i + Z_{i+1}$, the channel from X_i to $\hat{Y}_{i+1}$ is effectively an AWGN channel of noise power $D\sigma^2$ and gain h_i. Then the first term in the achievable rate expression becomes $\log\left(1 + \frac{|h_i|^2 P}{D\sigma^2}\right)$ which is greater than or equal to $\log\left(1 + \frac{|h_i|^2 P}{\sigma^2}\right) - \log(D)$.

Due to the coarse quantization, the second term in the achievable rate expression is reduced significantly as compared to quantizing at the noise level. We have

$$\begin{aligned} I(Y_{\mathcal{V}^i}; \hat{Y}_{\mathcal{V}^i} | X_{\mathcal{N}}, \hat{Y}_{\mathcal{N} \setminus \mathcal{V}^i}) &= I(Z_{\mathcal{V}^i}; \{Z + \hat{Z}\}_{\mathcal{V}^i}) \\ &= (|\mathcal{V}^i| - 1) \log\left(1 + \frac{\sigma^2}{(D-1)\sigma^2}\right) \\ &= i \log\left(1 + \frac{1}{D-1}\right) \\ &\leq \frac{i}{D-1} \\ &\leq 1, \end{aligned}$$

since $i \leq D - 1$. Since the capacity of the line network is given by the minimum of the capacities of each link: $\min_i \log(1 + |h_i|^2 P)$, we see that decreasing the resolution of quantization as the number of nodes increases results in a gap of $\log(D) + 1$ to capacity. If the quantization were done at the noise level, the first term in the noisy network coding achievable rate would suffer from only a $\log(2)$ decrease instead of $\log(D)$ with respect to capacity, however the second term would be linear in D, overall resulting in a gap to capacity that is linear in D.

At a first glance, coarser quantization resulting in better achievable rates might seem counter-intuitive. We discuss this in more depth in the following section.

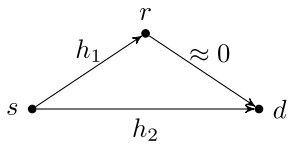

Fig. 5. Example.

V. Gap to Capacity With Noisy Network Coding

In this section, we discuss the elements of the gap between the rate achieved by noisy network coding (NNC) and the cutset bound and identify a trade-off between different elements of the gap. Our main result builds on the understanding of this trade-off.

Consider an arbitrary discrete memoryless network with a set of nodes $\mathcal{N}$ where a source node s wants to communicate to a destination node d with the help of the remaining nodes acting as relays. As stated earlier in Section II-D, noisy network coding can achieve the rate given in (3). Comparing this with the cutset bound on the capacity of the network,

$$\overline{C} = \sup_{p(x_{\mathcal{N}})} \min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; Y_{\Omega^c} \mid X_{\Omega^c}), \tag{10}$$

we observe the following differences. First, while the maximization in (10) is over all possible input distributions, only independent input distributions are admissible in (3). This gap corresponds to a potential beamforming gain that is allowed in the cutset bound but not exploited by NNC. Second, the first term in (3) is similar to (10) but with Y_{Ω^c} in (10) replaced by $\hat{Y}_{\Omega^c}$ in (3). This difference corresponds to a rate loss due to the quantization noise introduced by the relays. Third, there is the extra term $I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c})$ reducing the rate in (3). One way to potentially interpret this term would be as the rate penalty for communicating the quantized (compressed) observations $\hat{Y}_\Omega$ to the destination on top of the desired message. Note that this is the rate required to describe the observations Y_Ω at the distortion dictated by $\hat{Y}_\Omega$ to a decoder that already knows (or has decoded) $X_{\mathcal{N}}, \hat{Y}_{\Omega^c}$.

However, it is not completely clear if this interpretation is precise because the non-unique decoder employed by NNC does not require the quantization indices to be explicitly decoded. The non-unique decoder of NNC searches for the unique source codeword that is jointly typical with some (not necessarily unique) set of quantization indices at the relays and the received signal at the destination. The following example in Figure 5 illustrates that in certain cases the decoder can indeed recover the transmitted message even if it can not uniquely recover the quantization index of the relay. Even though we focus on the extremal case where the $r-d$ link is zero, the discussion extends to the case where this link is sufficiently weak.

Consider the classical relay channel with a very weak link from the relay to the destination. Clearly, as long as the source uses a codebook of rate less than the capacity of the direct link, no matter what the operation at the relay is, the destination can always decode the source message by performing a joint typicality test between its received signal and the source codebook (which is subsumed by the non-unique typicality test of NNC). In particular, if the relay quantizes too finely, then there is no way for the destination to recover the relay's quantization index, even though the source message can still be recovered.

On the other hand, this example reveals the following strange property of the expression in (3). While the above discussion reveals that in the setup of Fig. 5, the rate achieved by NNC is equal to the capacity of the direct link independent of the relay's operation (i.e. what $\hat{Y}_r$ is), the rate in (3) is decreasing with increasing resolution for the quantization at the relay (due to the subtractive term $I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c})$). This suggests a more careful analysis of the rate achieved by NNC which leads to the improved rate given in (11), as shown at the bottom of this page. Here, only those relays that are in $\mathcal{M} \subseteq \mathcal{N}$ are considered in the non-unique typicality decoding, while the other relay transmissions are treated as noise. For example, for the relay channel in Figure 5, this would correspond to not considering the relay in the typicality decoding.

It has been shown in [18] that if $\mathcal{M}^*$ is the subset that maximizes (11) for a given $\prod_{i\in\mathcal{N}} p(x_i)p(\hat{y}_i|y_i, x_i)$, then the quantization indices of the relays in $\mathcal{M}^*$ can be uniquely decoded at the destination, while the quantization indices of the relays in $\mathcal{N} \setminus \mathcal{M}^*$ cannot be decoded and in fact, it is optimal to treat the transmissions from these relays as noise. Since the transmissions from $\mathcal{N} \setminus \mathcal{M}^*$ are treated as noise, the expression (11) is increased if these relays are shut down. Hence, we can conclude that in the optimal distribution $\prod_{i\in\mathcal{N}} p(x_i)p(\hat{y}_i|y_i, x_i)$ for NNC, some relays can be off (not utilized or equivalently always quantizing their received signals to zero) and some relays can be active, but the quantization indices of all relays (the active ones and trivially the inactive ones) can be uniquely decoded at the destination. Since the quantization indices are communicated to the destination together with the source message, there should be a rate penalty for communicating them which is precisely the term $I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{M}}, \hat{Y}_{\Omega^c})$.

The above discussion reveals that NNC communicates not only the source message but also the quantization indices to the destination despite the non-unique typicality test performed at the decoder; and while making quantizations finer introduces less quantization noise in the communication, it leads to a larger rate penalty for communicating these quantization indices to the destination. This tradeoff is made explicit in Theorem 1 which establishes the following

$$\sup_{\prod_{i\in\mathcal{N}} p(x_i)p(\hat{y}_i|y_i,x_i)} \sup_{\mathcal{M}\subseteq\mathcal{N}} \min_{\Omega\subseteq\mathcal{M}: s\in\Omega, d\in\mathcal{M}\setminus\Omega} \Big(I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) - I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{M}}, \hat{Y}_{\Omega^c}) \Big) \tag{11}$$

achievable rate

$$\overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - \frac{N}{Q} - d_Q^* \log(Q+1),$$

for any $Q \geq 0$. Here, the term $\frac{N}{Q}$ corresponds to the rate penalty associated with communicating the quantization indices and the term $d_Q^* \log(Q+1)$ corresponds to the rate penalty due to the quantization noise. Choosing a larger Q increases the latter but decreases the former.

VI. Proof of Main Result

In this section we prove Theorem 1 by evaluating the rate achieved by noisy network coding in (3) for a specific choice of the distribution $\prod_{k\in\mathcal{N}} p(x_k)p(\hat{y}_k|y_k, x_k)$ that satisfies the power constraint. We choose the channel input vector at each node j as $\mathbf{X}_j \sim \mathcal{CN}(0, PI)$ and $\hat{Y}_k$ for each receive antenna in the network is chosen such that

$$\hat{Y}_k = Y_k + \hat{Z}_k \quad \text{where } \hat{Z}_k \sim \mathcal{CN}(0, Q\sigma^2), \tag{12}$$

independent of everything else, for some $Q \geq 0$. Then, the achievable rate stated in (3) is given by

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} \Big(I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) - I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) \Big). \tag{13}$$

This implies that the following rate is also achievable:

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) - \max_{\Omega: s\in\Omega, d\in\Omega^c} I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}). \tag{14}$$

We first show that for the choice of the distribution for X_j's and $\hat{Y}_k$'s in (12), we have $I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) \leq \frac{N}{Q}$ for all cuts Ω such that $s \in \Omega, d \in \Omega^c$, as follows.

$$\begin{aligned}
&I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) \\
&\quad = h(\hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) - h(\hat{Y}_\Omega|Y_\Omega, X_\mathcal{N}, \hat{Y}_{\Omega^c}) \\
&\quad \overset{(a)}{=} h(\hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) - h(\hat{Y}_\Omega|Y_\Omega, X_\mathcal{N}) \\
&\quad \leq h(\hat{Y}_\Omega|X_\mathcal{N}) - h(\hat{Y}_\Omega|Y_\Omega, X_\mathcal{N}) \\
&\quad \overset{(b)}{=} \left(\sum_{j\in\Omega} N_j\right) \log(Q+1) - \left(\sum_{j\in\Omega} N_j\right) \log(Q) \\
&\quad = \left(\sum_{j\in\Omega} N_j\right) \log\left(1 + \frac{1}{Q}\right) \\
&\quad \leq \frac{N}{Q},
\end{aligned} \tag{15}$$

where both (a) and (b) follow due to our specific choice for the distribution $\prod_{k\in\mathcal{N}} p(x_k)p(\hat{y}_k|y_k, x_k)$. Hence,

$$\max_{\Omega: s\in\Omega, d\in\Omega^c} I(Y_\Omega; \hat{Y}_\Omega|X_\mathcal{N}, \hat{Y}_{\Omega^c}) \leq \frac{N}{Q}. \tag{16}$$

We now lower bound the first term in (14). Since X_Ω is chosen to be $\mathcal{CN}(0, PI)$, the quantity $I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c})$ is equal to $C_Q^{i.i.d.}(\Omega)$, where $C_Q^{i.i.d.}(\Omega)$ is defined in (4). Let Ω_Q^* denote the cut with minimal cut value as defined in (5). Then,

$$\begin{aligned}
&\min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) \\
&\quad = \min_{\Omega: s\in\Omega, d\in\Omega^c} C_Q^{i.i.d.}(\Omega) \\
&\quad = C_Q^{i.i.d.}(\Omega_Q^*) \\
&\quad \overset{(a)}{\geq} C_0^{i.i.d.}(\Omega_Q^*) - d_Q^* \log(Q+1) \qquad (17) \\
&\quad \overset{(b)}{\geq} C_0^{i.i.d.}(\Omega_0^*) - d_Q^* \log(Q+1) \\
&\quad \overset{(c)}{\geq} \sup_{p(x_\mathcal{N})} I(X_{\Omega_0^*}; Y_{(\Omega_0^*)^c} \mid X_{(\Omega_0^*)^c}) \\
&\qquad - d_0^* \log\left(1 + \frac{\sum_{i\in\Omega_0^*} M_i}{d_0^*}\right) - d_Q^* \log(Q+1) \qquad (18) \\
&\quad \geq \sup_{p(x_\mathcal{N})} I(X_{\Omega_0^*}; Y_{(\Omega_0^*)^c} \mid X_{(\Omega_0^*)^c}) - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) \\
&\qquad - d_Q^* \log(Q+1) \\
&\quad = \sup_{p(x_\mathcal{N})} \min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; Y_{\Omega^c} \mid X_{\Omega^c}) - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) \\
&\qquad - d_Q^* \log(Q+1) \\
&\quad = \overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - d_Q^* \log(Q+1), \qquad (19)
\end{aligned}$$

where (a) is justified by the following:

$$\begin{aligned}
&C_Q^{i.i.d.}(\Omega_Q^*) \\
&\quad = \log\det\left(I + \frac{P}{(Q+1)\sigma^2} \mathbf{H}_{\Omega_Q^*\to(\Omega_Q^*)^c} \mathbf{H}^\dagger_{\Omega_Q^*\to(\Omega_Q^*)^c}\right) \\
&\quad \geq \log\det\left(I + \frac{P}{\sigma^2} \mathbf{H}_{\Omega_Q^*\to(\Omega_Q^*)^c} \mathbf{H}^\dagger_{\Omega_Q^*\to(\Omega_Q^*)^c}\right) \\
&\qquad - d_Q^* \log(Q+1) \\
&\quad = C_0^{i.i.d.}(\Omega_Q^*) - d_Q^* \log(Q+1),
\end{aligned} \tag{20}$$

(b) follows by the definition of Ω_0^* and (c) follows from [5, Lemma 6.6] equation (144), which considers a MIMO channel with per-antenna power constraint and bounds the gap between its capacity and the largest achievable rate with no spatial coding, i.e. the rate achieved by using independent inputs at the antennas.

The proof of Theorem 1 follows from (16) and (19). ∎

We next state an observation which will be useful in Section VIII when we analyze the static layered network.

Remark 1: If there exists a set of cuts $\mathcal{A}$ such that

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} C_Q^{i.i.d.}(\Omega) \geq \min_{\substack{\Omega\in\mathcal{A}: \\ s\in\Omega, d\in\Omega^c}} C_Q^{i.i.d.}(\Omega) - \kappa$$

for all Q, where κ is a constant, then the gap between the upper and the lower bound in Theorem 1 can be potentially improved to

$$\tilde{d}_0^* \log\left(1 + \frac{M}{\tilde{d}_0^*}\right) + \frac{N}{Q} + \tilde{d}_Q^* \log(Q+1) + \kappa, \tag{21}$$

$$R_{\text{NNC}} = \sup_{\prod_{k\in\mathcal{N}} p(x_k)p(\hat{y}_k|y_k,x_k)} \min_{\Omega: s\in\Omega, d\in\Omega^c} \Big(I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}, H) - I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c}, H) \Big), \tag{24}$$

where

$$\tilde{d}_Q^* \triangleq \text{DOF}\left(\underset{\substack{\Omega\in\mathcal{A}:\\ s\in\Omega, d\in\Omega^c}}{\arg\min}\ C_Q^{i.i.d.}(\Omega) \right). \tag{22}$$

This can be seen by modifying the proof of the lower bound (19) slightly as:

$$\begin{aligned}
&\min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}) \\
&\quad = \min_{\Omega: s\in\Omega, d\in\Omega^c} C_Q^{i.i.d.}(\Omega) \\
&\quad \geq \min_{\substack{\Omega\in\mathcal{A}:\\ s\in\Omega, d\in\Omega^c}} C_Q^{i.i.d.}(\Omega) - \kappa \\
&\quad \geq \min_{\substack{\Omega\in\mathcal{A}:\\ s\in\Omega, d\in\Omega^c}} C_0^{i.i.d.}(\Omega) - \tilde{d}_Q^* \log(Q+1) - \kappa \\
&\quad \geq \overline{C} - \tilde{d}_0^* \log\left(1 + \frac{M}{\tilde{d}_0^*}\right) - \tilde{d}_Q^* \log(Q+1) - \kappa,
\end{aligned}$$

where each step follows by the same arguments in (19).

VII. Fast-Fading Layered Network

In this section, we concentrate on the fast-fading layered network defined in Section II-B and obtain an approximation for the capacity of this network.

A. Applying Theorem 1 to the Fast-Fading Layered Network

For the fast-fading setup, we assume that the destination knows all the instantaneous channel realizations in the network while the source and the relay nodes only know the statistics of the channel coefficients. We first note that under this assumption, the cutset bound and the noisy network coding rate can be expressed as follows.

- *Cutset Bound:* Noting that under the above assumption the effective received signal at the destination can be considered to be (Y_d, H), where H contains all the channel realizations in the network, the cutset bound in (1) can be written as

$$\overline{C} = \sup_{p(x_{\mathcal{N}})} \Big(\min_{\Omega: s\in\Omega, d\in\Omega^c} \overline{C}(\Omega) \Big), \tag{23}$$

where

$$\begin{aligned}
\overline{C}(\Omega) &\triangleq I(X_\Omega; Y_{\Omega^c}, H|X_{\Omega^c}) \\
&= I(X_\Omega; Y_{\Omega^c}|X_{\Omega^c}, H)
\end{aligned}$$

since $X_{\mathcal{N}}$ is independent of H.

- *Noisy Network Coding:* The rate achieved by noisy network coding is given by (24), as shown at the top of this page, where we have again used the fact that $X_{\mathcal{N}}$ is independent of H.

We now proceed to the proof of Theorem 2. We first note that by following similar steps as in the proof of Theorem 1, we can get the following result:

$$\overline{C} \geq C \geq \overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - \frac{N}{Q} - d_Q^* \log(Q+1), \tag{25}$$

where d_Q^* is now analogously defined as the expected degrees of freedom of the fast-fading MIMO channel corresponding to the cut $\mathbf{\Omega}_Q^*$ that minimizes $\mathbb{E}[C_Q^{i.i.d.}(\mathbf{\Omega})]$, which we express as

$$d_Q^* \triangleq \text{DOF}\left(\underset{\Omega: s\in\Omega, d\in\Omega^c}{\arg\min}\ \mathbb{E}\left[C_Q^{i.i.d.}(\mathbf{\Omega}) \right] \right),$$

and the expectation is with respect to the randomness in the channels. Note that when we proved Theorem 1, we defined $C_Q^{i.i.d.}(\mathbf{\Omega})$ to be the first mutual information term in the achievable rate for noisy network coding in (13) when the input distributions $\mathbf{X}_j$ are i.i.d. $\mathcal{CN}(0, PI)$ and $\hat{Y}_k$'s are chosen according to (12). In the current fast-fading case the first mutual information term in the achievable rate for noisy network coding in (24) is equal to $\mathbb{E}[C_Q^{i.i.d.}(\mathbf{\Omega})]$ under the same distribution for the $\mathbf{X}_j$'s and $\hat{Y}_k$'s. Therefore, the proof of Theorem 1 can be applied verbatim in the current case by only modifying the definition of d_Q^* accordingly.

Now, by choosing Q to be equal to $Q' = D - 1$, we get that

$$\begin{aligned}
C &\geq \overline{C} - d_0^* \log\left(1 + \frac{M}{d_0^*}\right) - \frac{N}{Q'} - d_{Q'}^* \log(Q'+1) \\
&= \overline{C} - d_0^* \log\left(1 + \frac{K(D-1)}{d_0^*}\right) - \frac{K(D-1)}{Q'} \\
&\quad - d_{Q'}^* \log(Q'+1) \\
&\overset{(a)}{=} \overline{C} - K \log\left(1 + \frac{K(D-1)}{K}\right) - \frac{K(D-1)}{Q'} \\
&\quad - K \log(Q'+1) \\
&\overset{(b)}{\geq} \overline{C} - K \log D - K - K \log D, \\
&= \overline{C} - 2K \log D - K,
\end{aligned}$$

where

- (a) follows from Lemma 1, provided below, which states that $d_Q^* = K$ for any $Q \geq 0$; and
- (b) follows since $Q' = D - 1$.

Thus, we have characterized the capacity of the fast-fading layered network within a gap of $2K \log D + K$. The next subsection describes how this result can be tightened to obtain a gap equal to $K \log D + K$, which will conclude the proof of Theorem 2.

Lemma 1: For the fast-fading layered network, we have for any $Q \geq 0$,

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_Q^{i.i.d.}(\mathbf{\Omega}) \right] = \mathbb{E}\left[C_Q^{i.i.d.}(\mathcal{V}^0) \right],$$

which implies

$$d_Q^* = K.$$

Proof: See Appendix A. ∎

B. Tightening the Approximation

The main idea in tightening the approximation is that for the fast-fading layered network, we can get rid of the term $d_0^* \log\left(1+\frac{M}{d_0^*}\right)$ in the gap given by Theorem 1.

Recall from the proof of Theorem 1 that this term appears because we need to bound the difference between the capacity of a MIMO channel with per-antenna power constraint and the rate achievable by using independent inputs at each antenna. However, for an i.i.d. Rayleigh fast-fading MIMO channel, it is the case that independent inputs at each node are optimal and so the largest rate achievable by using independent inputs at each antenna is equal to the capacity [19].

Then, the proof for obtaining equation (25) which is based on the proof of Theorem 1 can be repeated verbatim except for one change: in (18), the term $d_0^* \log\left(1+\frac{\sum_{i\in\Omega_0^*} M_i}{d_0^*}\right)$ can be removed. This is valid since $\Omega_0^* = \mathcal{V}^0$ as shown by Lemma 1, which induces an i.i.d. Rayleigh fast-fading $K \times K$ MIMO channel. This improves the lower bound obtained in the previous subsection from $\overline{C} - 2K\log D - K$ to $\overline{C} - K\log D - K$. For clarity, we present the arguments in full formality below.

We first define, for any $Q \geq 0$,

$$f_Q(x,y) \triangleq \mathbb{E}\left[\log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{x,y}\mathbf{H}_{x,y}^\dagger\right)\right], \quad (26)$$

where $\mathbf{H}_{x,y}$ is a $x \times y$ matrix containing i.i.d. $\mathcal{CN}(0,1)$ entries. Note that using this notation, we have that $\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right]$ is equal to $f_Q(K,K)$.

Using this notation, the statement of Lemma 1 is

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right] = \mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right] = f_Q(K,K). \quad (27)$$

Before proceeding to the proof of the lower bound, we give the following lemma, which states that the cutset bound defined in (23), which involves a maximization over all possible input distributions, is equal to $\min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_0^{i.i.d.\cdot}(\Omega)\right]$.

Lemma 2: For the fast-fading layered network,

$$\overline{C} = \min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_0^{i.i.d.\cdot}(\Omega)\right],$$

and hence $\overline{C}$ *also equals* $\mathbb{E}\left[C_0^{i.i.d.\cdot}(\mathcal{V}^0)\right] = f_0(K,K)$.

Proof: See Appendix B. ∎

Using the above lemma, we can now complete the proof of the tighter lower bound via the following chain of inequalities. Recall that $\mathbf{X}_j$ are chosen to be i.i.d. $\mathcal{CN}(0, PI)$ and $\hat{Y}_k$'s are chosen according to (12). As in the previous subsection, we set Q to be equal to $Q' = D-1$.

$$\begin{aligned}
C &\overset{(a)}{\geq} \min_{\Omega: s\in\Omega, d\in\Omega^c} \Big(I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}, H) \\
&\qquad - I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c}, H)\Big) \\
&\geq \min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}, H) \\
&\quad - \max_{\Omega: s\in\Omega, d\in\Omega^c} I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c}, H) \\
&\overset{(b)}{\geq} \min_{\Omega: s\in\Omega, d\in\Omega^c} I(X_\Omega; \hat{Y}_{\Omega^c}|X_{\Omega^c}, H) - \frac{K(D-1)}{Q'} \\
&= \min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_{Q'}^{i.i.d.\cdot}(\Omega)\right] - \frac{K(D-1)}{Q'} \\
&\overset{(c)}{=} f_{Q'}(K,K) - \frac{K(D-1)}{Q'} \\
&\overset{(d)}{\geq} f_0(K,K) - K\log(Q'+1) - \frac{K(D-1)}{Q'} \\
&\overset{(e)}{=} \overline{C} - K\log(Q'+1) - \frac{K(D-1)}{Q'} \\
&= \overline{C} - K\log D - K, \quad (28)
\end{aligned}$$

where

- (a) gives the rate achieved by noisy network coding,
- (b) follows since, similar to (15),

$$\max_{\Omega: s\in\Omega, d\in\Omega^c} I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, \hat{Y}_{\Omega^c}, H) \leq \frac{K(D-1)}{Q'},$$

- (c) follows from (27),
- (d) follows, similarly to (17), because

$$\begin{aligned}
&f_{Q'}(K,K) \\
&= \mathbb{E}\left[\log\det\left(I + \frac{P}{(Q'+1)\sigma^2}\mathbf{H}_{K,K}\mathbf{H}_{K,K}^\dagger\right)\right] \\
&\geq \mathbb{E}\left[\log\det\left(I + \frac{P}{\sigma^2}\mathbf{H}_{K,K}\mathbf{H}_{K,K}^\dagger\right)\right] \\
&\quad - K\log(Q'+1) \\
&= f_0(K,K) - K\log(Q'+1), \quad (29)
\end{aligned}$$

- (e) follows from Lemma 2. Note the difference between this step and the corresponding step (18) in the proof of Theorem 1. For general networks, the term $d_0^* \log\left(1+\frac{\sum_{i\in\mathcal{N}} M_i}{d_0^*}\right)$ is required, while for the special case of fast-fading layered networks, we are able to get rid of it.

This concludes the proof of Theorem 2. ∎

C. Proof of Corollary 1

In this subsection, we prove that the result of Theorem 2 can be extended to the case with multiple sources. Assume that K single-antenna sources each wish to transmit a message at rate $\frac{R}{K}$, so that the sum-rate is R. We have, via the cutset bound, the following upper bound on the achievable sum-rate K:

$$R < \sup_{p(x_{\mathcal{N}})} \min_{\Omega:\, s_1,s_2,\ldots,s_K\in\Omega,\, d\in\Omega^c} I(X_\Omega; Y_{\Omega^c}|X_{\Omega^c}, H).$$

The RHS of the above expression is equal to the cutset bound on the achievable rate in the case of a single source

as given in (23). Hence, we have that if a sum-rate R is achievable, then it must satisfy

$$R < \overline{C}.$$

This proves the upper bound on the sum-capacity. In the remainder of this subsection, we focus on proving the lower bound. As before, we fix the distribution $p(x_{\mathcal{N}})$ to be $\prod_{k\in\mathcal{N}} p(x_k)$, with each term being $\mathcal{CN}(0, P)$. The distribution $p(\hat{y}_k|y_k, x_k)$ at the relays is to be of the same form as that in (12). From the result for multiple sources stated in [6, Th. 1], we get that R is achievable if for all $1 \le k \le K$, we have

$$k\frac{R}{K} < \min_{\substack{\Omega:|\{s_i:s_i\in\Omega\}|=k,\\ d\in\Omega^c}} \Big(I(X_\Omega; Y_{\Omega^c}|X_{\Omega^c}, H) - I(Y_\Omega; \hat{Y}_\Omega|X_{\mathcal{N}}, Y_{\Omega^c}, H)\Big). \quad (30)$$

For a given k, the above constraint is obtained by considering cuts Ω which contain k source nodes and therefore it upper bounds the sum rate kR/K achievable for these k sources.

Note that we get a constraint on R for each value of k, where $k \in \{1, 2, \ldots, K\}$. Also, note that if we consider $k = K$, we get a constraint on R that is the same as (24). So, if this were the only constraint on R, then the proof of Theorem 2 in Section VII-B, which shows that the right-hand side of (24) is larger than $\overline{C} - K\log D - K$, would conclude the proof of Corollary 1. Towards this goal, we prove in Appendix C that any $k < K$ imposes a constraint on R that is only looser than the constraint

$$\begin{aligned} R &< \overline{C} - K\log D - K \\ &= f_0(K, K) - K\log D - K. \end{aligned}$$

This concludes the proof of Corollary 1. ∎

VIII. Static Layered Networks

In this section, we prove Theorem 3. We first show that for any $Q \ge 0$, $\min_{\Omega:s\in\Omega,d\in\Omega^c} C_Q^{i.i.d.\cdot}(\Omega)$ can be approximated upto an additive constant by restricting the minimization to cuts in a particular class. Then, Theorem 3 is proved by making use of Remark 1.

For convenience, let $\mathbf{H}_{\mathcal{V}_i\to\mathcal{V}_{i+1}}$ denote the matrix in $\mathbb{C}^{K\times K}$ containing channel gains from nodes in layer i to nodes in layer $i+1$, and call the K^2 entries in $\mathbf{H}_{\mathcal{V}_i\to\mathcal{V}_{i+1}}$ as the links in layer i. With this convention in mind, let $\mathcal{A}$ denote the set of cuts Ω for which the links crossing from Ω to Ω^c come from at most $K-1$ layers, e.g. see Figure 6.

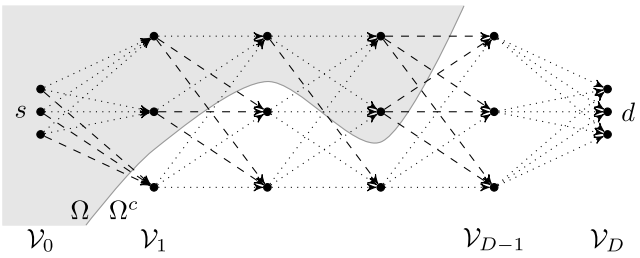

Fig. 6. The cut Ω depicted here $\notin \mathcal{A}$ since the crossing links come from 4 layers, and $4 > K - 1 = 2$.

Lemma 3: For the static layered network in Section II-C, we have, for any $Q \ge 0$,

$$\min_{\Omega:s\in\Omega,d\in\Omega^c} C_Q^{i.i.d.\cdot}(\Omega) \le \min_{\substack{\Omega\in\mathcal{A}:\\ s\in\Omega,d\in\Omega^c}} C_Q^{i.i.d.\cdot}(\Omega),$$

and

$$\min_{\Omega:s\in\Omega,d\in\Omega^c} C_Q^{i.i.d.\cdot}(\Omega) \ge \min_{\substack{\Omega\in\mathcal{A}:\\ s\in\Omega,d\in\Omega^c}} C_Q^{i.i.d.\cdot}(\Omega) - K\log K.$$

Proof: The upper bound is immediate. The lower bound can be proved by noting that the chain of inequalities given on bottom of this page, holds for any cut $\Omega \notin \mathcal{A}$, where (a) follows since for any cut $\notin \mathcal{A}$, at least K terms in the summation are non-zero and each of these terms can be lower-bounded by the AWGN capacity of a point-to-point channel between a single transmit and single receive antenna with unit magnitude channel coefficient; and (b) follows by Lemma 4 which is stated and proved below. This concludes the proof of the lemma. ∎

Lemma 4: For the static layered network in Section II-C, we have, for any $Q \ge 0$,

$$C_Q^{i.i.d.\cdot}(\mathcal{V}_0) \le K\log\left(1 + \frac{P}{(Q+1)\sigma^2}\right) + K\log K.$$

Proof:

$$\begin{aligned} C_Q^{i.i.d.\cdot}(\mathcal{V}_0) &= \log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{\mathcal{V}_0\to\mathcal{V}_1}\mathbf{H}_{\mathcal{V}_0\to\mathcal{V}_1}^\dagger\right) \\ &\overset{(a)}{\le} \sum_{i=1}^{K}\log\left(1 + \frac{P}{(Q+1)\sigma^2}\mathbf{h}_i\mathbf{h}_i^\dagger\right) \\ &\overset{(b)}{=} \sum_{i=1}^{K}\log\left(1 + \frac{P}{(Q+1)\sigma^2}K\right) \\ &\le K\log\left(1 + \frac{P}{(Q+1)\sigma^2}\right) + K\log K, \end{aligned}$$

$$\begin{aligned} C_Q^{i.i.d.\cdot}(\Omega) &= \sum_{i=0}^{D-1}\log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{(\mathcal{V}_i\cap\Omega)\to(\mathcal{V}_{i+1}\cap\Omega^c)}\mathbf{H}_{(\mathcal{V}_i\cap\Omega)\to(\mathcal{V}_{i+1}\cap\Omega^c)}^\dagger\right) \\ &\overset{(a)}{\ge} K\log\left(1 + \frac{P}{(Q+1)\sigma^2}\right) \\ &\overset{(b)}{\ge} C_Q^{i.i.d.\cdot}(\mathcal{V}_0) - K\log K \\ &\ge \min_{\substack{\Omega\in\mathcal{A}:\\ s\in\Omega,d\in\Omega^c}} C_Q^{i.i.d.\cdot}(\Omega) - K\log K \end{aligned}$$

where $\mathbf{h}_i$ denotes the ith row of $H_{\mathcal{V}_0\to\mathcal{V}_1}$ and (a) follows by using Hadamard's inequality and (b) follows from the fact that the channel gains have unit magnitude. ∎

We now use the observation made in Remark 1 to prove Theorem 3. As in the previous section, first note that $M = N = K(D-1)$. Then, we note that for any cut Ω in $\mathcal{A}$, the matrix $\mathbf{H}_{\Omega\to\Omega^c}$ can have at most $K(K-1)$ columns. This is because the links crossing from Ω to Ω^c come from at most $K-1$ layers, hence there can be at most $K(K-1)$ nodes in Ω from which the crossing links originate. Hence, a trivial upper bound on $\tilde{d}^*_Q$ (defined in (22)) for any Q is

$$\tilde{d}^*_Q \le K(K-1) \le K^2. \tag{31}$$

Now, we set Q to be $Q' = D-1$ and use the result in (21) to prove Theorem 3 as follows:

$$\begin{aligned}
C &\ge \overline{C} - \tilde{d}^*_0 \log\left(1+\frac{M}{\tilde{d}^*_0}\right) - \frac{N}{Q'} - \tilde{d}^*_{Q'}\log(Q'+1) - \kappa \\
&\overset{(a)}{\ge} \overline{C} - \tilde{d}^*_0 \log\left(1+\frac{M}{\tilde{d}^*_0}\right) - \frac{N}{Q'} - \tilde{d}^*_{Q'}\log(Q'+1) \\
&\quad - K\log K \\
&\overset{(b)}{\ge} \overline{C} - K^2\log\left(1+\frac{K(D-1)}{K^2}\right) - \frac{K(D-1)}{Q'} \\
&\quad - K^2\log(Q'+1) - K\log K \\
&\overset{(c)}{=} \overline{C} - K^2\log\left(1+\frac{D-1}{K}\right) - K - K^2\log D \\
&\quad - K\log K \\
&\ge \overline{C} - 2K^2\log D - K\log K - K,
\end{aligned}$$

where (a) follows by Lemma 3, (b) follows from (31) and the fact that $x\log(1+M/x)$ is an increasing function of x, and (c) follows since $Q' = D-1$. This concludes the proof of Theorem 3. ∎

IX. Concluding Remarks

In this paper, we have developed improved capacity approximations for Gaussian relay networks. While existing approximations bound the capacity gap only in terms of the total number of nodes in the network, we have developed a refined approximation for the capacity of general Gaussian relay networks where the gap depends not only on the total number of nodes but other structural properties of the network (the degrees of freedom of the mincut). We have shown that this refined result allows to better approximate the capacity of many Gaussian networks, some classes of layered networks in particular.

The improvement comes from carefully exploiting a trade-off inherent to compress-and-forward based strategies. When relays quantize/compress signals very finely, little quantization noise is introduced to the communication. When relays quantize/compress signals coarsely, there is a smaller rate penalty associated with communicating these quantization indices to the destination. We have shown that this trade-off can be very much in favor of coarse quantization, leading to the counter-intuitive principle of quantizing signals more and more coarsely with increasing number of relaying stages.

Appendix A
Proof of Lemma 1

Proof: By the definition of $C_Q^{i.i.d.\cdot}(\Omega)$,

$$\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right] = \mathbb{E}\left[\log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{\Omega\to\Omega^c}\mathbf{H}^\dagger_{\Omega\to\Omega^c}\right)\right].$$

We first note that for any cut Ω in the set $\{\mathcal{V}^0, \mathcal{V}^1, \ldots, \mathcal{V}^{D-1}\}$, the statistics of $\mathbf{H}_{\Omega\to\Omega^c}$ are identical. Hence, the value of $\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right]$ is the same for all these cuts and we use $\mathcal{V}^0$ as a representative.

We now prove the statement: For any $Q \ge 0$,

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right] = \mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right]. \tag{32}$$

The proof of the "$\le$" direction of the inequality, i.e.

$$\min_{\Omega: s\in\Omega, d\in\Omega^c} \mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right] \le \mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right]$$

is immediate. We focus on proving the inequality in the other direction in the remainder of this proof.

Consider a cut Ω that contains M_1 nodes from $\mathcal{V}_1$, M_2 from $\mathcal{V}_2$ and so on until M_{D-1} from $\mathcal{V}_{D-1}$ (see Figure 6). Then $\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right]$ is given by

$$\mathbb{E}\left[\log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{\Omega\to\Omega^c}\mathbf{H}^\dagger_{\Omega\to\Omega^c}\right)\right],$$

where $\mathbf{H}_{\Omega\to\Omega^c}$ is a block diagonal matrix containing blocks of size M_1^c-by-K, M_2^c-by-M_1, M_3^c-by-M_2, …, M_{D-1}^c-by-M_{D-2} and finally K-by-M_{D-1}. In the preceding sentence, we have abused notation slightly by using M_i^c to mean $|\mathcal{V}_i| - M_i = K - M_i$.

Since $\mathbf{H}_{\Omega\to\Omega^c}$ has a block diagonal structure, $\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right]$ breaks down into a sum of terms, each being a function of the number of nodes in Ω that belong to two adjacent layers. Thus,

$$\begin{aligned}
&\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\Omega)\right] \\
&\quad= \mathbb{E}\left[\log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{\Omega\to\Omega^c}\mathbf{H}^\dagger_{\Omega\to\Omega^c}\right)\right] \\
&\quad= f_Q(M_1^c, K) + f_Q(M_2^c, M_1) \\
&\qquad + \cdots + f_Q(M_{D-1}^c, M_{D-2}) + f_Q(K, M_{D-1}),
\end{aligned} \tag{33}$$

where $f_Q(x,y)$ is defined as in (26):

$$f_Q(x,y) \triangleq \mathbb{E}\left[\log\det\left(I + \frac{P}{(Q+1)\sigma^2}\mathbf{H}_{x,y}\mathbf{H}^\dagger_{x,y}\right)\right],$$

and $\mathbf{H}_{x,y}$ is a $x\times y$ matrix containing i.i.d. $\mathcal{CN}(0,1)$ entries. Note that using this notation, $\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right]$ is equal to $f_Q(K,K)$. So, our aim is to show that for any cut Ω, the quantity appearing in (33) is no less than $f_Q(K,K)$.

To accomplish this, we note the following properties of the function $f_Q(x,y)$:

a) $f_Q(x,y) = f_Q(y,x)$.
b) $f_Q(z,y) \ge f_Q(x,y)$ if $z \ge x$.
c) $f_Q(x,y) + f_Q(K-x,y) \ge f_Q(K,y)$.

The first two properties are straightforward and the third property follows via a simple application of Hadamard's inequality.

Proving that the quantity in (33) is no less than $f_Q(K,K)$ is just a matter of applying these properties multiple times. For concreteness, we show this for the case $D=4$ below, which can be generalized in a straightforward fashion to higher values of D.

$$\begin{aligned}
&f_Q(M_1^c,K)+f_Q(M_2^c,M_1)+f_Q(M_3^c,M_2)+f_Q(K,M_3)\\
&\quad\geq f_Q(M_1^c,K)+f_Q(M_2^c,M_1)\\
&\qquad+f_Q(M_3^c,M_2)+f_Q(M_2,M_3)\\
&\quad\geq f_Q(M_1^c,K)+f_Q(M_2^c,M_1)+f_Q(K,M_2)\\
&\quad\geq f_Q(M_1^c,K)+f_Q(M_2^c,M_1)+f_Q(M_1,M_2)\\
&\quad\geq f_Q(M_1^c,K)+f_Q(K,M_1)\\
&\quad\geq f_Q(K,K)\\
&\quad=\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right], \qquad (34)
\end{aligned}$$

where the first inequality follows by applying property (b) to the last term in the first line, the second inequality follows by applying (c) to the last two terms in the earlier line etc. Since this is true for any cut $\mathbf{\Omega}$, we have shown that

$$\min_{\Omega:s\in\Omega,d\in\Omega^c}\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathbf{\Omega})\right]\geq\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right]. \qquad (35)$$

Thus, we have shown that (32) is true, i.e.

$$\min_{\Omega:s\in\Omega,d\in\Omega^c}\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathbf{\Omega})\right]=\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathcal{V}^0)\right]=f_Q(K,K), \qquad (36)$$

which implies that $\mathcal{V}^0\in\arg\min_{\Omega:s\in\Omega,d\in\Omega^c}\mathbb{E}\left[C_Q^{i.i.d.\cdot}(\mathbf{\Omega})\right]$. This further implies that

$$d_Q^*=K,$$

since the DOF of the fast-fading MIMO channel corresponding to $\mathcal{V}^0$ is K. ∎

Appendix B
Proof of Lemma 2

Starting from (23), we have

$$\begin{aligned}
\overline{C}&=\sup_{p(x_{\mathcal{N}})}\left(\min_{\Omega:s\in\Omega,d\in\Omega^c}\overline{C}(\mathbf{\Omega})\right)\\
&=\sup_{p(x_{\mathcal{N}})}\left(\min_{\Omega:s\in\Omega,d\in\Omega^c}I(X_\Omega;Y_{\Omega^c}|X_{\Omega^c},H)\right)\\
&\leq\sup_{p(x_{\mathcal{N}})}\left(I(X_{\mathcal{V}^0};Y_{\mathcal{V}^0}|X_{(\mathcal{V}^0)^c},H)\right)\\
&\overset{(a)}{=}\mathbb{E}\left[\log\det\left(I+\frac{P}{\sigma^2}\mathbf{H}_{\mathcal{V}^0\to(\mathcal{V}^0)^c}\mathbf{H}^{\dagger}_{\mathcal{V}^0\to(\mathcal{V}^0)^c}\right)\right]\\
&=\mathbb{E}\left[C_0^{i.i.d.\cdot}(\mathcal{V}^0)\right]\\
&\overset{(b)}{=}\min_{\Omega:s\in\Omega,d\in\Omega^c}\mathbb{E}\left[C_0^{i.i.d.\cdot}(\mathbf{\Omega})\right]\\
&\leq\sup_{p(x_{\mathcal{N}})}\left(\min_{\Omega:s\in\Omega,d\in\Omega^c}I(X_\Omega;Y_{\Omega^c}|X_{\Omega^c},H)\right)\\
&=\overline{C},
\end{aligned}$$

where (a) follows by the fact that for a i.i.d. Rayleigh fast-fading MIMO channel, the optimal input distribution is independent across antennas [19], and (b) follows from (32) which shows that the cut that minimizes $\mathbb{E}\left[C_0^{i.i.d.\cdot}(\mathbf{\Omega})\right]$ is $\mathcal{V}^0$. ∎

Appendix C

In this appendix, we elaborate on the argument required to prove the lower bound in Corollary 1.

Consider a cut $\mathbf{\Omega}$ such that $|\{s_i:s_i\in\mathbf{\Omega}\}|=k$. Let $\mathbf{\Omega}$ contain M_i nodes from layer $\mathcal{V}_i$, for $1\leq i\leq D-1$. As before, we choose the quantization noise variance Q to be $Q'=D-1$. This gives us a constraint on the achievable sum-rate R as follows:

$$\begin{aligned}
R&<\frac{K}{k}\left(I(X_\Omega;\hat{Y}_{\Omega^c}|X_{\Omega^c},H)-I(Y_\Omega;\hat{Y}_\Omega|X_{\mathcal{N}},\hat{Y}_{\Omega^c},H)\right)\\
&=\frac{K}{k}\left(\mathbb{E}\left[C_{Q'}^{i.i.d.\cdot}(\mathbf{\Omega})\right]-I(Y_\Omega;\hat{Y}_\Omega|X_{\mathcal{N}},\hat{Y}_{\Omega^c},H)\right)\\
&=\frac{K}{k}\Big(f_{Q'}(M_1^c,k)+f_{Q'}(M_2^c,M_1)+\cdots+f_{Q'}(K,M_{D-1})\\
&\qquad-I(Y_\Omega;\hat{Y}_\Omega|X_{\mathcal{N}},\hat{Y}_{\Omega^c},H)\Big),
\end{aligned}$$

where we use the notation $f_Q(x,y)$ defined in (26). Since we have

$$I(Y_\Omega;\hat{Y}_\Omega|X_{\mathcal{N}},\hat{Y}_{\Omega^c},H)\leq\frac{\sum_{i=1}^{D-1}M_i}{Q'}=\frac{\sum_{i=1}^{D-1}M_i}{D-1},$$

which can be proved using steps similar to those used to arrive at (15), we can impose a tighter constraint on the sum-rate R due to the cut $\mathbf{\Omega}$, which is as follows.

$$\begin{aligned}
R<\frac{K}{k}\Bigg(&f_{Q'}(M_1^c,k)+f_{Q'}(M_2^c,M_1)\\
&+\cdots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\Bigg). \qquad (37)
\end{aligned}$$

In the following, we show for any $k<K$, the above is weaker than

$$R<f_0(K,K)-K\log D-K, \qquad (38)$$

i.e. the right-hand side of (37) for any $k<K$ is larger than $f_0(K,K)-K\log D-K$.

Note that if $f_0(K,K)-K\log D-K\leq 0$, the achievable rate claimed by (38) is zero so there is nothing to prove, so we assume that $f_0(K,K)-K\log D-K>0$.

- If the cut $\mathbf{\Omega}$ has $M_1=M_2=\cdots=M_{D-1}=0$, then the expression in the constraint (37) becomes

$$\begin{aligned}
&\frac{K}{k}\Bigg(f_{Q'}(M_1^c,k)+f_{Q'}(M_2^c,M_1)\\
&\qquad+\cdots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\Bigg)\\
&=\frac{K}{k}f_{Q'}(K,k)\\
&\overset{(a)}{\geq}f_{Q'}(K,K)\\
&\geq f_{Q'}(K,K)-K\\
&\overset{(b)}{\geq}f_0(K,K)-K\log D-K,
\end{aligned}$$

$$\frac{1}{k\binom{K}{k}}\sum_{1\le i_1<\cdots<i_k\le K}\log\det\left(\pi e\left(I_k+\lambda\,\mathbf{H}^{\dagger}_{l,(i_1,\ldots,i_k)}\mathbf{H}_{l,(i_1,\ldots,i_k)}\right)\right)\;\ge\;\frac{1}{K}\log\det\left(\pi e\left(I_K+\lambda\,\mathbf{H}^{\dagger}_{l,K}\mathbf{H}_{l,K}\right)\right)$$

where (a) follows from Claim 1, provided at the end of this Appendix, and (b) follows by the same argument as in (29).

- If the cut is $\mathbf{\Omega}$ such that $M_i = K$ for some $i \in \{1, 2, \ldots, K\}$, then

$$\begin{aligned}
&\frac{K}{k}\left(f_{Q'}(M_1^c,k)+f_{Q'}(M_2^c,M_1)\right.\\
&\qquad\left.+\cdots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\right)\\
&\overset{(a)}{\ge}\frac{K}{k}\left(f_{Q'}(K,K)-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\right)\\
&\ge\frac{K}{k}\left(f_{Q'}(K,K)-K\right)\\
&\overset{(b)}{\ge} f_{Q'}(K,K)-K\\
&\ge f_0(K,K)-K\log D-K,
\end{aligned}$$

where (a) follows by using the properties of the function f_Q as in (34), and (b) follows since $\frac{K}{k}\ge 1$.

- Let $i^* = \arg\max_{1\le i\le D-1} M_i$ so that $M_{i^*} = \max_{1\le i\le D-1} M_i$. From the previous two cases, we can focus our attention to $0 < M_{i^*} < K$. Also, note that $M_1 < K$ implies that $M_1^c > 0$. The RHS of the constraint due to $\mathbf{\Omega}$ is

$$\begin{aligned}
&\frac{K}{k}\left(f_{Q'}(M_1^c,k)+f_{Q'}(M_2^c,M_1)\right.\\
&\qquad\left.+\cdots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\right)\\
&=\frac{K}{k}f_{Q'}(M_1^c,k)+\frac{K}{k}\\
&\quad\times\left(f_{Q'}(M_2^c,M_1)+\ldots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\right)\\
&\overset{(a)}{\ge} f_{Q'}(M_1^c,K)+\frac{K}{k}\\
&\quad\times\left(f_{Q'}(M_2^c,M_1)+\ldots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\right)\\
&\overset{(b)}{\ge} f_{Q'}(M_1^c,K)\\
&\quad+\left(f_{Q'}(M_2^c,M_1)+\ldots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\right)\\
&\overset{(c)}{\ge} f_{Q'}(K,K)-K\\
&\ge f_0(K,K)-K\log D-K,
\end{aligned}$$

where

- (a) follows by Claim 1,
- (b) follows because $\frac{K}{k}\ge 1$ and because

$$f_{Q'}(M_2^c,M_1)+\cdots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1},$$

is non-negative, which is proved as follows:

$$\begin{aligned}
&f_{Q'}(M_2^c,M_1)+\cdots+f_{Q'}(K,M_{D-1})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\\
&\quad\ge f_{Q'}(K,M_{i^*})-\frac{\sum_{i=1}^{D-1}M_i}{D-1}\\
&\quad\ge f_{Q'}(K,M_{i^*})-M_{i^*}\\
&\quad\ge\frac{M_{i^*}}{K}f_{Q'}(K,K)-M_{i^*}\\
&\quad=\frac{M_{i^*}}{K}\left(f_{Q'}(K,K)-K\right)\\
&\quad\ge\frac{M_{i^*}}{K}\left(f_0(K,K)-K\log D-K\right)\\
&\quad\ge 0,
\end{aligned}$$

- (c) follows by noting that the expression in (b) is the constraint on sum-rate imposed by a cut which is $\mathcal{V}_0\cup\mathbf{\Omega}$, which we know is lower bounded by $f_{Q'}(K,K)-K$.

The above analysis shows that (38) renders all other constraints redundant. ∎

Claim 1: For any $Q\ge 0$, any $k\in\{1,2,\ldots,K-1\}$ and any $l\in\{1,2,\ldots,K\}$,

$$\frac{K}{k}f_Q(l,k)\ge f_Q(l,K).$$

Proof: Recall that $f_Q(l,K)$ is defined to be

$$\mathbb{E}\left[\log\det\left(I+\frac{P}{(Q+1)\sigma^2}\mathbf{H}^{\dagger}_{l,K}\mathbf{H}_{l,K}\right)\right].$$

To be more explicit in the following, we write I_p to denote an identity matrix of size p. Also, for brevity, we denote $\frac{P}{(Q+1)\sigma^2}$ by λ. For any fixed $\mathbf{H}_{l,K}$, we have by [20, eq. (3.15)] the inequality given at the top of this page, where $\mathbf{H}_{l,(i_1,\ldots,i_k)}$ is obtained by choosing the columns of $\mathbf{H}_{l,K}$ indexed by $(i_1,\ldots,i_k)$.

Hence,

$$\begin{aligned}
&\frac{1}{k\binom{K}{k}}\sum_{1\le i_1<\cdots<i_k\le K}\log\det\left(I_k+\lambda\,\mathbf{H}^{\dagger}_{l,(i_1,\ldots,i_k)}\mathbf{H}_{l,(i_1,\ldots,i_k)}\right)\\
&\qquad+\frac{1}{k}\log\left((\pi e)^k\right)\\
&\quad\ge\frac{1}{K}\log\left((\pi e)^K\right)+\frac{1}{K}\log\det\left(I_K+\lambda\,\mathbf{H}^{\dagger}_{l,K}\mathbf{H}_{l,K}\right),
\end{aligned}$$

which means

$$\begin{aligned}
&\frac{1}{k\binom{K}{k}}\sum_{1\le i_1<\cdots<i_k\le K}\log\det\left(I+\lambda\,\mathbf{H}^{\dagger}_{l,(i_1,\ldots,i_k)}\mathbf{H}_{l,(i_1,\ldots,i_k)}\right)\\
&\qquad\ge\frac{1}{K}\log\det\left(I+\lambda\,\mathbf{H}^{\dagger}_{l,K}\mathbf{H}_{l,K}\right).
\end{aligned}$$

Now, taking expectation on both sides and observing that each term in the summation has identical statistics, the desired claim is proved. ∎

REFERENCES

[1] R. Kolte and A. Özgür, "Improved capacity approximations for Gaussian relay networks," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.

[2] R. Kolte, A. Özgür, and A. El Gamal, "Optimized noisy network coding for Gaussian relay networks," in *Proc. IEEE Int. Zürich Seminar Commun.*, Feb. 2014, pp. 140–143.

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

[4] G. Kramer, I. Marić, and R. D. Yates, "Cooperative communications," *Found. Trends Network.*, vol. 1, nos. 3–4, pp. 271–425, 2007.

[5] A. Avestimehr, S. Diggavi, and D. Tse, "Wireless network information flow: A deterministic approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1872–1905, Apr. 2011.

[6] S. H. Lim, Y.-H. Kim, A. El Gamal, and S.-Y. Chung, "Noisy network coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 3132–3152, May 2011.

[7] A. Özgür and S. N. Diggavi, "Approximately achieving Gaussian relay network capacity with lattice-based QMF codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8275–8294, Dec. 2013.

[8] A. Raja and P. Viswanath, "Compress-and-forward scheme for relay networks: Backword decoding and connection to bisubmodular flows," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5627–5638, Sep. 2014.

[9] G. Kramer and J. Hou, "Short-message quantize-forward network coding," in *Proc. IEEE 8th Int. Workshop Multi-Carrier Syst. Solutions*, May 2011, pp. 1–3.

[10] S. H. Lim, K. T. Kim, and Y.-H. Kim, "Distributed decode-forward for multicast," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2014, pp. 636–640.

[11] B. Chern and A. Özgür, "Achieving the capacity of the N-relay Gaussian diamond network within log N bits," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7708–7718, Dec. 2014.

[12] A. Sengupta, I.-H. Wang, and C. Fragouli, "Optimizing quantize-map-and-forward relaying for Gaussian diamond networks," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2012, pp. 381–385.

[13] U. Niesen, B. Nazer, and P. Whiting, "Computation alignment: Capacity approximation without noise accumulation," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3811–3832, Jun. 2013.

[14] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[15] B. Nazer, M. Gastpar, S. A. Jafar, and S. Vishwanath, "Ergodic interference alignment," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6355–6371, Oct. 2012.

[16] T. Courtade and A. Özgür, "Approximate capacity of Gaussian relay networks: Is a sublinear gap to the cutset bound plausible?" in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 2251–2255, paper Th-PM2-2.3.

[17] A. El Gamal, "On information flow in relay networks," in *Proc. Nat. Telecommun. Conf. (NTC)*, vol. 2. Dec. 1981, pp. D4.1.1–D4.1.4.

[18] X. Wu and L.-L. Xie, "On the optimal compressions in the compress-and-forward relay schemes," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2613–2628, May 2013.

[19] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Dec. 1999.

[20] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inf. Control*, vol. 36, no. 2, pp. 133–156, Feb. 1978.

Ritesh Kolte (S'13) is a doctoral candidate in Electrical Engineering at Stanford University. His research interests are in the general area of information theory, optimization, statistical learning and inference, signal processing. He received the Bachelor of Technology and Master of Technology degrees from the Indian Institute of Technology Bombay in 2010 and the Master of Science degree from Stanford University in 2012, all in Electrical Engineering; and is a recipient of the Stanford Graduate Fellowship.

Ayfer Özgür (M'06) received her B.Sc. degrees in electrical engineering and physics from Middle East Technical University, Turkey, in 2001 and the M.Sc. degree in communications from the same university in 2004. From 2001 to 2004, she worked as hardware engineer for the Defense Industries Development Institute in Turkey. She received her Ph.D. degree in 2009 from the Information Processing Group at EPFL, Switzerland. In 2010 and 2011, she was a post-doctoral scholar with the Algorithmic Research in Network Information Group at EPFL. She is currently an Assistant Professor in the Electrical Engineering Department at Stanford University. Her research interests include network communications, wireless systems, and information and coding theory. Dr. Özgür received the EPFL Best Ph.D. Thesis Award in 2010 and a NSF CAREER award in 2013.

Abbas El Gamal (S'71–M'73–SM'83–F'00) is the Hitachi America Professor in the School of Engineering and the Chair of the Department of Electrical Engineering at Stanford University. He received his B.Sc. Honors degree from Cairo University in 1972, and his M.S. in Statistics and Ph.D. in Electrical Engineering both from Stanford University in 1977 and 1978, respectively. From 1978 to 1980, he was an Assistant Professor of Electrical Engineering at USC. From 2004 to 2009, he was the Director of the Information Systems Laboratory at Stanford University. His research contributions have been in network information theory, FPGAs, and digital imaging devices and systems. He has authored or coauthored over 220 papers and holds 35 patents in these areas. He is coauthor of the book *Network Information Theory* (Cambridge Press 2011). He received several honors and awards for his research contributions, including the 2012 Claude E. Shannon Award and the 2004 INFOCOM Paper Award. He is a member of the US National Academy of Engineering and a Fellow of the IEEE. He has been serving on the Board of Governors of the Information Theory Society since 2009 and is currently its Junior Past President. He has played key roles in several semiconductor, EDA, and biotechnology startup companies.