



HAL
open science

On the Multiplexing Gain of Discrete-Time MIMO Phase Noise Channels

Sheng Yang, Shlomo Shamai Shitz

► **To cite this version:**

Sheng Yang, Shlomo Shamai Shitz. On the Multiplexing Gain of Discrete-Time MIMO Phase Noise Channels. *IEEE Transactions on Information Theory*, 2017, 63 (4), pp.2394-2408. 10.1109/TIT.2017.2653718 . hal-01551485

HAL Id: hal-01551485

<https://centralesupelec.hal.science/hal-01551485>

Submitted on 28 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the Multiplexing Gain of Discrete-Time MIMO Phase Noise Channels

Sheng Yang, *Member, IEEE*, Shlomo Shamai (Shitz), *Fellow, IEEE*

Abstract—The capacity of a point-to-point discrete-time multi-input-multiple-output (MIMO) channel with phase uncertainty (MIMO phase noise channel) is still open. As a matter of fact, even the pre-log (multiplexing gain) of the capacity in the high signal-to-noise ratio (SNR) regime is unknown in general. We make some progresses in this direction for two classes of such channels. With phase noise on the individual paths of the channel (model A), we show that the multiplexing gain is $\frac{1}{2}$, which implies that the capacity *does not* scale with the channel dimension at high SNR. With phase noise at both the input and output of the channel (model B), the multiplexing gain is upper-bounded by $\frac{1}{2} \min\{n_t, (n_r - 2)^+ + 1\}$, and lower-bounded by $\frac{1}{2} \min\{n_t, \lfloor \frac{n_r + 1}{2} \rfloor\}$, where n_t and n_r are the number of transmit and receive antennas, respectively. The multiplexing gain is enhanced to $\frac{1}{2} \min\{n_t, n_r\}$ without receive phase noise, and to $\frac{1}{2} \min\{2n_t - 1, n_r\}$ without transmit phase noise. In all the cases of model B, the multiplexing gain scales linearly with $\min\{n_t, n_r\}$. Our main results rely on the derivation of non-trivial upper and lower bounds on the capacity of such channels.

Index Terms—Phase noise channel, multiple-input-multiple-output (MIMO), channel capacity, duality upper bound, multiplexing gain.

I. INTRODUCTION

The capacity of a point-to-point multiple-input-multiple-output (MIMO) Gaussian channel is well known in the coherent case, i.e., when the channel state information is available at the receiver [1], [2]. The capacity of the noncoherent MIMO channels, however, is still open in general. Nevertheless, asymptotic results of such channels, e.g., at high signal-to-noise ratio (SNR), have been obtained in some important cases.

In the seminal paper [3], Lapidot and Moser proposed a powerful technique, called the duality approach, that can be applied to a large class of fading channels and derived the exact high SNR capacity up to an $o(1)$ term. In particular, when the differential entropy of the channel matrix is finite, i.e., $h(\mathbf{H}) > -\infty$, it was shown in [3] that the pre-log (a.k.a. multiplexing gain), of the capacity is 0 and the high-SNR capacity is $\log \log \text{SNR} + \chi(\mathbf{H}) + o(1)$ where $\chi(\mathbf{H})$ is the so-called fading number of the channel. In addition, capacity upper and lower bounds for the MIMO Rayleigh and Ricean channels were obtained and shown to be tight at both low and high SNR regimes. In [4], Zheng and Tse showed that for noncoherent block fading MIMO Rayleigh channels with

with coherence time T , the pre-log is $M^*(1 - M^*/T)$ where $M^* \triangleq \min\{n_t, n_r, \lfloor \frac{T}{2} \rfloor\}$ with n_t and n_r being the number of transmit and receive antennas, respectively. In this work, we are interested in the MIMO phase noise channels in which the phases of the channel coefficients are not perfectly known.

Applying the duality approach and the “escape-to-infinity” property of the channel input, Lapidot characterized the high-SNR capacity of the discrete-time phase noise channel in the single-antenna case [5]. It was shown in [6] that the capacity-achieving input distribution is in fact discrete. Recently, capacity upper and lower bounds of the single-antenna channels with Wiener phase noise have been extensively studied in the context of optical fiber and microwave communications (see [7], [8], [9] and the references therein). In these works, the upper bounds are derived via duality and lower bounds are computed numerically using the auxiliary channel technique proposed in [10]. In particular, in [9], Durisi *et al.* investigated the MIMO phase noise channel with a common phase noise, a scenario motivated by the microwave link with centralized oscillators. The SIMO and MISO channels with common and separate phase noises are considered in [11]. The 2×2 MIMO phase noise channel with independent transmit and receive phase noises at each antenna was studied in [12], where the authors showed that the multiplexing gain is $\frac{1}{2}$ for a specific class of input distributions. For general MIMO channels with separate phase noises, estimation and detection algorithms have been proposed in [13], [14]. However for such channels, even the multiplexing gain is unknown, to the best of our knowledge.

In this work, we make some progresses in this direction. We consider two classes of discrete-time stationary and ergodic MIMO phase noise channels: model A with individual phase noises on the entries of the channel matrix, and model B with individual phase noises at the input and the output of the channel instead. The phase noise processes in both models are assumed to have finite differential entropy rate. For model A, we obtain the exact multiplexing gain $\frac{1}{2}$ for any channel dimension, which implies that the capacity *does not* scale with the channel dimension at high SNR. For model B with both transmit and receive phase noises, we show that the multiplexing gain is upper-bounded by $\frac{1}{2} \min\{n_t, (n_r - 2)^+ + 1\}$, and lower-bounded by $\frac{1}{2} \min\{n_t, \lfloor \frac{n_r + 1}{2} \rfloor\}$. The upper and lower bounds coincide for $n_r \leq 3$ or $n_r \geq 2n_t - 1$. Further, when receive phase noise is absent, the multiplexing gain is improved and we obtain the exact value of $\frac{1}{2} \min\{n_t, n_r\}$. If the transmit phase noise is absent instead, the multiplexing gain becomes $\frac{1}{2} \min\{2n_t - 1, n_r\}$.

The main technical contribution of this paper is two-fold. First, we derive a non-trivial upper bound on the capacity of

S. Yang is with LSS, CentraleSupélec, 3 rue Joliot-Curie, 91190 Gif-sur-Yvette, France. (e-mail: sheng.yang@centralesupelec.fr)

S. Shamai (Shitz) is with Technion-Israel Institute of Technology, Haifa, Israel. (e-mail: sshlomo@ee.technion.ac.il)

The work of S. Shamai was supported by the Israel Science Foundation (ISF). The material in this paper was presented in part at the 2016 IEEE Information Theory Workshop.

the MIMO phase noise channel with separate phase noises. The novelty of the upper bound lies in the finding of a suitable auxiliary distributions with which we apply the duality upper bound [15], [16], [3]. It is worth mentioning that, the class of single-variate Gamma output distributions, as the essential ingredient that led to the tight capacity upper bounds on previously studied channels, are not suitable for MIMO phase noise channels in general. In this paper, we introduce a class of *multi*-variate Gamma distributions that, combined with the duality upper bound, allows us to obtain a complete pre-log characterization for model A and partially for model B. The second contribution is the derivation of the capacity lower bounds for model B, based on the remarkable property of the differential entropy of the output vector in this channel. Namely, we prove that, at high SNR, the pre-log of the said entropy can go beyond the rank of the channel matrix, $\min\{n_t, n_r\}$, and scales as $n_r \log \text{SNR}$ as long as $n_r \leq 2n_t - 1$. The upper and lower bounds suggest that, with $n_r \geq 2n_t - 1$ receive antennas, n_t transmitted real symbols can be recovered at high SNR. This result has an interesting interpretation based on dimension counting. Let us consider the example of independent and memoryless transmit and receive phase noises uniformly distributed in $[0, 2\pi)$. In this case, phases of the input and the output do not contain any useful information, only the amplitudes matter. Note that the n_r output amplitudes are (non-linear) equations of $2n_t - 1$ unknowns, namely, the n_t input amplitudes and the $n_t - 1$ relative input phases, assuming the additive noises are negligible at high SNR. It is now not too hard to believe that with $n_r = 2n_t - 1$ equations, the receiver can successfully decode the n_t input amplitudes by solving the equations. This is however not possible with $n_r < 2n_t - 1$, in which case there are too many unknowns as compared to the number of equations. Nonetheless, we can reduce the number of active transmit antennas to $n'_t < n_t$ such that $2n'_t - 1 \leq n_r$, which means that the achievable multiplexing gain is $\frac{n'_t}{2} \leq \frac{1}{2} \lfloor \frac{n_t + 1}{2} \rfloor$. A formal proof in Section VI validates such an argument.

The remainder of the paper is organized as follows. The system model and main results are presented in Section II. Some preliminaries useful for the proof of the main results are provided in Section III. The upper bounds are derived in Section IV and Section V. We prove the lower bound for model B in section VI. Concluding remarks are given in Section VII. Most of the proofs are presented in the main body of the paper, with some details deferred to the Appendix.

II. SYSTEM MODEL AND MAIN RESULTS

Notation

Throughout the paper, we use the following notational conventions. For random quantities, we use upper case letters, e.g., X , for scalars, upper case letters with bold and non-italic fonts, e.g., \mathbf{V} , for vectors, and upper case letter with bold and sans serif fonts, e.g., \mathbf{M} , for matrices. Deterministic quantities are denoted in a rather conventional way with italic letters, e.g., a scalar x , a vector \mathbf{v} , and a matrix \mathbf{M} . Logarithms are in base 2. The Euclidean norm of a vector and a matrix is denoted by $\|\mathbf{v}\|$ and $\|\mathbf{M}\|$, respectively. The transpose and

conjugated transpose of \mathbf{M} are \mathbf{M}^\top and \mathbf{M}^H , respectively. \mathbf{H}^\dagger is the pseudo-inverse of a *tall* matrix \mathbf{H} . The argument (phase) of a complex value x is denoted by $\angle x \in [0, 2\pi)$. We use $\mathbf{A} \circ \mathbf{B}$ to denote the Hadamard (point-wise) product between vectors/matrices. x_{n+1}^{n+k} is a k -tuple or a column vector of $(x_{n+1}, \dots, x_{n+k})$; for brevity sometimes x^k replaces x_1^k . For convenience, wherever confusion is improbable, elementary scalar functions applied to a vector, e.g., $|\mathbf{x}|$ or $\cos(\boldsymbol{\theta})$, stand for a point-wise map on each element of the vector, and return a vector with the same dimension as the argument. We use $(\theta)_{2\pi}$ to denote $(\theta \bmod 2\pi)$, and $(x)^+ = \max\{x, 0\}$. $\Gamma(x)$ is the gamma function. We also use c_0 to represent a bounded constant whose value is irrelevant but may change at each occurrence. Similarly, c_H is a constant that may depend on \mathbf{H} but the value is irrelevant and bounded for almost all \mathbf{H} .

A. Channel model

In this paper, we are interested in a class of discrete-time MIMO phase noise channels with n_t transmit antennas and n_r receive antennas, defined by

$$\mathbf{Y}_t = (\mathbf{H} \circ e^{j\boldsymbol{\Theta}_t}) \mathbf{x}_t + \mathbf{Z}_t, \quad t = 1, 2, \dots, N,$$

where the deterministic channel matrix \mathbf{H} belongs to a set $\mathcal{H} \subset \mathbb{C}^{n_r \times n_t}$ of generic matrices¹; $\mathbf{x}_t \in \mathbb{C}^{n_t \times 1}$ is the input vector at time t , with the average power constraint $\frac{1}{N} \sum_{t=1}^N \|\mathbf{x}_t\|^2 \leq P$; the additive noise process $\{\mathbf{Z}_t\}$ is assumed to be spatially and temporally white with $\mathbf{Z}_t \sim \mathcal{CN}(0, \mathbf{I}_{n_r})$; $\boldsymbol{\Theta}_t$ is the matrix of phase noises on the individual entries of \mathbf{H} at time t ; the phase noise process $\{\boldsymbol{\Theta}_t\}$ is stationary and ergodic, and is independent of the additive noise process $\{\mathbf{Z}_t\}$. Both $\{\mathbf{Z}_t\}$ and $\{\boldsymbol{\Theta}_t\}$ are unknown to the transmitter and the receiver. Since the additive noise power is normalized, the transmit power P is identified with the SNR throughout the paper. The end-to-end channel is captured by the random channel matrix $\mathbf{H} \triangleq [h_{ik} e^{\boldsymbol{\Theta}_{ik}}]_{i,k}$.

In this paper, we consider two types of discrete-time phase noise processes² according to the spatial structures, as shown in Fig. 1:

- Model A refers to channels with phase uncertainty on the individual paths (path phase noise), such that the sequence $\{\boldsymbol{\Theta}_t\}$ has finite entropy rate

$$h(\{\boldsymbol{\Theta}_t\}) > -\infty.$$

It corresponds to the case where the phase information of the channel cannot be obtained accurately, e.g., in optical fiber communications. This model covers the channel with spatially independent phase noises as a special case.

- Model B refers to channels with phase noises at the input and/or output, i.e., $\boldsymbol{\Theta}_{ik} = \boldsymbol{\Theta}_{R,i} + \boldsymbol{\Theta}_{T,k}$. The vector $\boldsymbol{\Theta}_T \triangleq [\boldsymbol{\Theta}_{T,i}]_{i=1}^{n_t}$ contains the n_t phase noises at the transmit antennas, and $\boldsymbol{\Theta}_R \triangleq [\boldsymbol{\Theta}_{R,k}]_{k=1}^{n_r}$ is the vector of the n_r

¹It means that the channel matrix \mathbf{H} does not lie on any algebraic hypersurface. If we draw the entries of \mathbf{H} i.i.d. from a continuous distribution, then \mathbf{H} is generic almost surely.

²The limitation of discrete-time phase noise model, which ignores the filtering before sampling in practical continuous-time communication systems, has been discussed in [17] and [8].

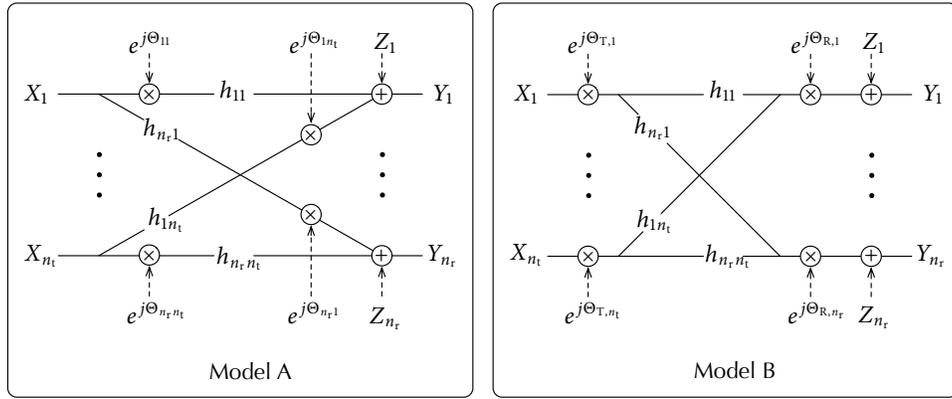


Fig. 1. Two models considered in this work: model A with path phase noise, and model B with transmit/receive phase noise.

phase noises at the receive antennas. This model captures the phase corruption at both the transmit and receive RF chains, e.g., caused by imperfect oscillators. We consider three cases of model B: leftmargin=.5in

- B1) with both transmit and receive phase noises such that $h(\{\Theta_{T,t}, \Theta_{R,t}\}) > -\infty$;
- B2) with only transmit phase noise such that $h(\{\Theta_{T,t}\}) > -\infty$;
- B3) with only receive phase noise such that $h(\{\Theta_{R,t}\}) > -\infty$.

Note that model B1 covers the case where both the transmitter and receiver use separate (and imperfect) oscillators for different antennas, whereas models B2 and B3 contain the case with centralized oscillators on one side and separate oscillators at the other side.

The capacity of such a stationary and ergodic channel is [3], [18]

$$C(P) \triangleq \lim_{N \rightarrow \infty} \sup \frac{1}{N} I(\mathbf{X}^N; \mathbf{Y}^N), \quad (1)$$

where the supremum is taken over all distributions with the average power constraint

$$\frac{1}{N} \sum_{k=1}^N \mathbb{E} [\|\mathbf{X}_k\|^2] \leq P.$$

Our work focuses on the multiplexing gain r of such a channel, defined as the pre-log of the capacity $C(P)$ as $P \rightarrow \infty$,

$$r \triangleq \lim_{P \rightarrow \infty} \frac{C(P)}{\log P}.$$

B. Main results

The main results of this work are summarized as follows, and are illustrated in Fig. 2. First, the case with common phase noise is rather straightforward from [9].

Proposition 1. *With common phase noise, i.e., $\Theta_t = \Theta_t \mathbf{1}_{n_r \times n_t}$ and $h(\{\Theta_t\}) > -\infty$, the multiplexing gain is $\min\{n_t, n_r\} - \frac{1}{2}$.*

Proof: The proof is provided in Appendix A. ■

Then our new results are on channels with separate phase noises either on the individual paths (model A) or at the input/output (model B) of the channel.

Theorem 1. *The multiplexing gain of model A is $\frac{1}{2}$.*

The above result shows that extra transmit and receive antennas do not improve the multiplexing gain of a channel with phase uncertainty on each path of the channel. The achievability of the single-antenna case was shown in [5]. Our main contribution lies in the converse, as will be shown in Section IV.

Theorem 2. *The multiplexing gain of model B is*

- upper-bounded by $\frac{1}{2} \min\{n_t, (n_r - 2)^+ + 1\}$, and lower-bounded by $\frac{1}{2} \min\{n_t, \lfloor \frac{n_r + 1}{2} \rfloor\}$ with both transmit and receive phase noises, the upper bound is achievable when $n_r \leq 3$ or $n_r \geq 2n_t - 1$;
- $\min\{\frac{n_r}{2}, \frac{n_t}{2}\}$, with only transmit phase noise;
- $\min\{\frac{n_r}{2}, n_t - \frac{1}{2}\}$, with only receive phase noise.

Interestingly, the multiplexing gain of model B depends on the number of transmit and receive antennas differently, which is rarely the case for previously studied point-to-point MIMO channels.

Remark II.1. *As shown in Fig. 2, transmit phase noise is more detrimental than receive phase noise, and strictly so when $n_r > n_t > 1$. Intuitively, with transmit phase noise each transmitted symbol is accompanied by a different phase noise symbol, which means that no more than half of the total spatial degrees of freedom is available for useful signal. On the other hand, with receive phase noise, although half of the received signal dimension is occupied by phase noises, it is enough to increase the number of receive antennas to recover almost all transmitted symbols.*

Remark II.2. *Obviously, the multiplexing gain of model B1 is upper-bounded by that of models B2 and B3. Such a “trivial” upper bound is given by $\min\{\frac{n_t}{2}, \frac{n_r}{2}, n_t - \frac{1}{2}\} = \min\{\frac{n_t}{2}, \frac{n_r}{2}\}$. When $n_r \leq n_t$, the optimal multiplexing gain is $\frac{n_r}{2}$ with phase noises at either side of the channel, whereas no more than $\frac{(n_r - 2)^+ + 1}{2}$ is achievable with phase noises at both sides. These are the cases for which model B1 is strictly “worse” than both models B2 and B3. When $n_r \geq 2n_t - 1$, with transmit phase*

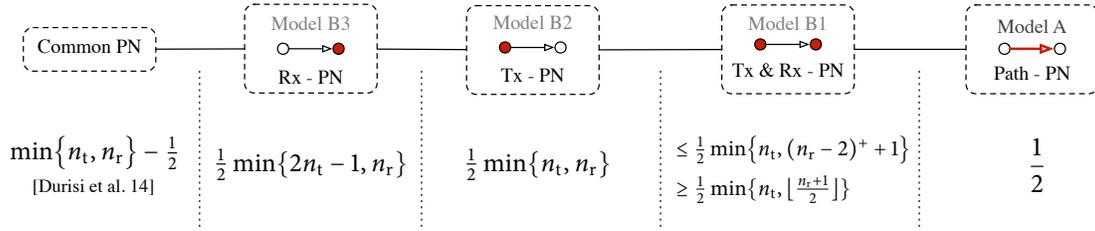


Fig. 2. Multiplexing gain of the MIMO phase noise channels.

noise, the optimal multiplexing gain is $\frac{n_t}{2}$ regardless of the presence of receive phase noise.

Remark II.3. Theorem 2 shows that, when $n_t = n_r = 2$ and 3, the exact multiplexing gain of model B1 is $\frac{(n_r-2)^++1}{2}$ which gives $\frac{1}{2}$ and 1, respectively. In contrast, the trivial upper bound provides 1 and $\frac{3}{2}$, respectively. These are the two cases of model B1 for which we obtain exact multiplexing gain that is strictly lower than that of models B2 and B3.

The remainder of the paper is dedicated to the proof of the main results. We start with some mathematical preliminaries.

III. PRELIMINARIES

Definition 1 (Multivariate Gamma distribution [19]). The n -variate Gamma distribution has the following density function

$$p(\mathbf{s}) = s_1^{\alpha_1-1} (s_2 - s_1)^{\alpha_2-1} \cdots (s_n - s_{n-1})^{\alpha_n-1} \cdot g_\alpha \mu^{\alpha_1+\cdots+\alpha_n} \exp(-\mu s_n),$$

for $0 < s_1 < s_2 < \cdots < s_n$, and 0 elsewhere, with $\mu > 0$ and $\alpha_i > 0$, $i = 1, \dots, n$; $g_\alpha \triangleq 1 / (\prod_{i=1}^n \Gamma(\alpha_i))$ is the normalization factor. When $n = 1$, we have the standard single-variate Gamma distribution,

$$p(s) = g_\alpha \mu^\alpha s^{\alpha-1} \exp(-\mu s), \quad s > 0, \alpha > 0. \quad (2)$$

Lemma 1 (Monotonicity of $\mathbb{E}_{\chi_k^2(\lambda)}[\log X]$ over λ). Let X be a non-central Chi-square distribution with degrees of freedom k and noncentrality parameter λ , denoted as $X \sim \chi_k^2(\lambda)$. Then the expected logarithm of X is strictly increasing with $\lambda \geq 0$, for any $k \in \mathbb{N}$.

Proof: The case with $k = 2n$ is known and has been proved in [3]. In the following, we provide a simple proof for the general case of k , although we are only interested in the case $k = 1$ later in the paper. Let us define $f_k(\lambda) \triangleq \mathbb{E}_{\chi_k^2(\lambda)}[\log X]$. The probability density function of $\chi_k^2(\lambda)$ is [20]

$$p_{\chi_k^2(\lambda)}(x) = \sum_{l=0}^{\infty} \frac{e^{-\frac{\lambda}{2}} \left(\frac{\lambda}{2}\right)^l}{l!} p_{\chi_{k+2l}^2(0)}(x), \quad x \geq 0.$$

Then it readily follows from the definition of $f_k(\lambda)$ that

$$f_k(\lambda) = \sum_{l=0}^{\infty} \frac{e^{-\frac{\lambda}{2}} \left(\frac{\lambda}{2}\right)^l}{l!} f_{k+2l}(0).$$

To prove that $f_k(\lambda)$ is increasing with λ , it is enough to show that the derivative of $f_k(\lambda)$ with respect to λ is positive. Indeed,

$$\begin{aligned} f'_k(\lambda) &= -\frac{1}{2} f_k(\lambda) + \frac{1}{2} \sum_{l=0}^{\infty} \frac{e^{-\frac{\lambda}{2}} \left(\frac{\lambda}{2}\right)^l}{l!} f_{k+2(l+1)}(0) \\ &= \frac{1}{2} (f_{k+2}(\lambda) - f_k(\lambda)) \\ &= \frac{1}{2} (\mathbb{E}[\log(X+Y)] - \mathbb{E}[\log X]) \\ &> 0, \end{aligned}$$

where we used the fact that if $X \sim \chi_k^2(\lambda)$ and $Y \sim \chi_2^2(0)$, then $X+Y \sim \chi_{k+2}^2(\lambda)$. ■

Lemma 2 (Change of variables [18]). Let $\mathbf{Y} = f(\mathbf{X})$ with a bijective map $f: \mathbb{R}^m \rightarrow \mathbb{R}^m$. Then

$$h(\mathbf{Y}) = h(\mathbf{X}) + \mathbb{E}[\log |\det(\mathbf{J})|],$$

where $\mathbf{J} \triangleq \left[\frac{\partial Y_k}{\partial X_l} \right]_{k,l=1,\dots,m}$ is the Jacobian matrix.

Lemma 3. If each element of the n -vector \mathbf{X} is circularly symmetric with independent phases, and the probability density function (pdf) of \mathbf{X} exists with respect to the Lebesgue measure, then

$$\begin{aligned} p_{|\mathbf{X}|}(|\mathbf{x}|) &= 2\pi \prod_{i=1}^n |x_i| p_X(\mathbf{x}) \\ p_{|\mathbf{X}|^2}(|\mathbf{x}|^2) &= \pi p_X(\mathbf{x}). \end{aligned} \quad (3)$$

Further, if $h(\mathbf{X}) > -\infty$, we have

$$\begin{aligned} h(\mathbf{X}) &= h(|\mathbf{X}|) + \sum_{i=1}^n \mathbb{E}[\log |X_i|] + n \log 2\pi \\ &= h(|\mathbf{X}|^2) + n \log \pi. \end{aligned}$$

Lemma 4. Let $\mathbf{X} \in \mathbb{C}^n$ with $h(\mathbf{X}) > -\infty$. Then

$$h(\mathbf{X}) = h(|\mathbf{X}|^2) + h(\angle \mathbf{X} | |\mathbf{X}|) - n. \quad (4)$$

Let $\Theta \in [0, 2\pi)^n$ be independent of \mathbf{X} and $h(\Theta) > -\infty$. Then

$$|h(e^{j\Theta} \circ \mathbf{X}) - h(|\mathbf{X}|^2)| \leq c_0. \quad (5)$$

Proof: Applying Lemma 2 twice, we readily obtain (4)

$$\begin{aligned} h(\mathbf{X}) &= h(|\mathbf{X}|, \angle \mathbf{X}) + \sum_{k=1}^n \mathbb{E} \log |X_k| \\ &= h(|\mathbf{X}|) + h(\angle \mathbf{X} | |\mathbf{X}|) + \sum_{k=1}^n \mathbb{E} \log |X_k| \\ &= h(|\mathbf{X}|^2) + h(\angle \mathbf{X} | |\mathbf{X}|) - n. \end{aligned}$$

To prove (5), we introduce Φ that is uniformly distributed in $[0, 2\pi)^n$ and independent of \mathbf{X} and Θ , then

$$\begin{aligned} h(e^{j\Theta} \circ \mathbf{X}) &= h(e^{j(\Theta+\Phi)} \circ \mathbf{X} | \Phi) \\ &\leq h(e^{j\Phi'} \circ \mathbf{X}) \\ &= h(|\mathbf{X}|^2) + n \log \pi, \end{aligned}$$

where $\Phi' \triangleq (\Theta + \Phi)_{2\pi}$ is uniformly distributed in $[0, 2\pi)^n$, and from (4),

$$\begin{aligned} h(e^{j\Theta} \circ \mathbf{X}) &= h(|\mathbf{X}|^2) + h((\angle \mathbf{X} + \Theta)_{2\pi} | |\mathbf{X}|) - n \\ &\geq h(|\mathbf{X}|^2) + h((\angle \mathbf{X} + \Theta)_{2\pi} | |\mathbf{X}|, \angle \mathbf{X}) - n \\ &= h(|\mathbf{X}|^2) + h(\Theta) - n. \end{aligned}$$

Hence, (5) holds with the constant c_0 corresponding to $\max\{|h(\Theta) - n|, n \log \pi\}$. ■

Lemma 5. For any $\Theta \in [0, 2\pi)$ with $h(\Theta) > -\infty$,

$$\mathbb{E}[\log |\sin(\Theta)|] > \frac{h(\Theta) - \log(B(\frac{1-\alpha}{2}, \frac{1}{2}))}{\alpha}, \quad \forall \alpha \in (0, 1), \quad (6)$$

where $B(x, y)$ is the Beta function. Thus, $\mathbb{E}[\log |\sin(\Theta)|] > -\infty$. Let $\Theta \in [0, 2\pi)^n$. If $h(\Theta) > -\infty$, then

$$h(\cos(\Theta)) > -\infty. \quad (7)$$

Proof: To prove (6), we introduce an auxiliary distribution with density $q(\theta) = \beta |\sin(\theta)|^{-\alpha}$, $\theta \in [0, 2\pi)$, with $\alpha \in (0, 1)$ and $\beta \triangleq \frac{1}{2B(\frac{1-\alpha}{2}, \frac{1}{2})}$. Then it follows that $h(\Theta) + \mathbb{E}[\log(q(\Theta))] = -D(p \| q) \leq 0$ where $D(\cdot \| \cdot)$ is the Kullback-Leibler divergence, which yields (6). We proceed to prove (7),

$$\begin{aligned} h(\cos(\Theta)) &\geq h(\cos(\Theta) | \Omega) \\ &= h(\Theta | \Omega) + \sum_{k=1}^n \mathbb{E}[\log |\sin(\Theta_k)|] \\ &= h(\Theta) - I(\Omega; \Theta) + \sum_{k=1}^n \mathbb{E}[\log |\sin(\Theta_k)|] \\ &> -\infty, \end{aligned}$$

where we partition $[0, 2\pi)^n$ in such a way that $\cos(\Theta)$ is a bijective function of Θ in each partition indexed by Ω ; the first equality is from Lemma 2; the last inequality is from the boundedness of $h(\Theta)$, the fact that Ω only takes a finite number of values, and the application of (6). ■

Lemma 6. Let $\mathbf{V} \in \mathbb{R}^{m \times 1}$ with $h(\mathbf{V}) > -\infty$ and $\mathbb{E}[\|\mathbf{V}\|^2] < \infty$. Then

$$\inf_{\mathbf{x} \in \mathbb{R}^m: \|\mathbf{x}\|=1} \mathbb{E}[\log |\mathbf{V}^\top \mathbf{x}|] > -\infty.$$

Proof: This is a straightforward adaptation of the result in [3, Lemma 6.7-f] for the complex case. The real case can be proved by following the same steps. To be self-contained, we provide an alternative proof as follows. Define $V_x \triangleq |\mathbf{V}^\top \mathbf{x}|^2$, and one can verify from the assumptions that $h(V_x) > -\infty$ and $\mathbb{E}[V_x] \leq \infty$. We introduce an auxiliary pdf $q(V_x)$ based on the Gamma distribution defined in (2) with some $\alpha \in (0, 1)$.

Then we have for any $V_x \in \mathbb{R}^+$, $h(V_x) \leq \mathbb{E}[-\log q(V_x)] = (1-\alpha)\mathbb{E}[\log V_x] + \mu\mathbb{E}[V_x] + c_0$ which yields $\mathbb{E}[\log V_x] \geq (1-\alpha)^{-1}(h(V_x) - \mu\mathbb{E}[V_x] - c_0) > -\infty$. ■

Lemma 7. Let $\Theta \in [0, 2\pi)$ with $h(\Theta) > -\infty$ and be independent of some $Z \sim \mathcal{CN}(0, 1)$, then for any given $\beta \in \mathbb{C}$,

$$|h(\beta e^{j\Theta} + Z) - \log^+ |\beta|| \leq c_0, \quad (8)$$

$$\text{and } |h(|\beta + Z|)| \leq c'_0. \quad (9)$$

Proof: First we prove (8). When $|\beta| \leq 1$, we have $\log^+ |\beta| = 0$. It follows that $h(\beta e^{j\Theta} + Z) \geq h(\beta e^{j\Theta} + Z | \Theta) = \log(\pi e)$ and $h(\beta e^{j\Theta} + Z) \leq \log(\pi e(\text{Var}(\beta e^{j\Theta} + Z))) \leq \log(2\pi e)$, which proves (8) for $|\beta| \leq 1$. Next, we assume that $|\beta| > 1$. It is without loss of generality to consider $\beta \in \mathbb{R}^+$. Let Z_R and Z_I be the real and imaginary parts of Z , respectively. Then

$$\begin{aligned} h(\beta e^{j\Theta} + Z) &= h(\beta \cos(\Theta) + Z_R) \\ &\quad + h(\beta \sin(\Theta) + Z_I | \beta \cos(\Theta) + Z_R) \\ &\geq h(\beta \cos(\Theta)) + h(Z_I) \\ &= \log \beta + h(\cos(\Theta)) + \frac{1}{2} \log(\pi e). \quad (10) \end{aligned}$$

Since $\beta e^{j\Theta} + Z = e^{j\Theta}(\beta + \tilde{Z})$ where $\tilde{Z} \triangleq Z e^{-j\Theta} \sim \mathcal{CN}(0, 1)$ is independent of Θ , we can apply (5) from Lemma 4,

$$\begin{aligned} h(\beta e^{j\Theta} + Z) &\leq h(\beta^2 + |Z|^2 + 2\beta|Z| \cos(\Theta - \angle Z)) + c_0 \\ &= h(|Z|^2 + 2\beta|Z| \cos(\Theta - \angle Z)) + c_0 \\ &\leq \frac{1}{2} \log(2\pi e \text{Var}(|Z|^2 + 2\beta|Z| \cos(\Theta - \angle Z))) + c_0 \\ &\leq \log \beta + \frac{1}{2} \log(2\pi e (\text{Var}(|Z|^2) + \text{Var}(2|Z|))) + c_0, \quad (11) \end{aligned}$$

where we use the condition $\beta > 1$. The lower bound (10) and upper bound (11) complete the proof of (8) for $|\beta| > 1$. To prove (9), we introduce some Θ uniformly distributed in $[0, 2\pi)$, then $h(|\beta + Z|) = h(\beta e^{j\Theta} + Z) = h(\beta e^{j\Theta} + Z) - \mathbb{E}[\log(|\beta + Z|)] - \log 2\pi$. It can be shown that $|\mathbb{E}[\log(|\beta + Z|)] - \log^+ |\beta|| \leq c_0$. To see this, we write $\mathbb{E}[\log(|\beta + Z|)] = \frac{1}{2} \mathbb{E}[\log(|Z|^2 + |\beta|^2 + 2|\beta Z| \cos(\angle \beta + Z))]$ where $\angle_{\beta+Z}$ is uniformly distributed in $[0, 2\pi)$ and independent of the other variables. Taking expectation over $\angle_{\beta+Z}$, we obtain $\mathbb{E}[\log(|\beta + Z|)] = \frac{1}{2} \mathbb{E}[\log(|Z|^2 + |\beta|^2)] + c_0$, since $\frac{1}{2\pi} \int_0^{2\pi} \log(a + b \cos \theta) d\theta = \log \frac{a + \sqrt{a^2 - b^2}}{2}$, for all $a \geq b > 0$. Then, applying Jensen's inequality with expectation over Z , we have $\mathbb{E}[\log(|\beta + Z|)] \leq \frac{1}{2} \log(1 + |\beta|^2) + c_0 \leq \log^+ |\beta| + c'_0$. Using the monotonicity of the logarithmic function, we also have $\mathbb{E}[\log(|\beta + Z|)] = \frac{1}{2} \mathbb{E}[\log(|Z|^2 + |\beta|^2)] + c_0 \geq \max\{\log |\beta|, \mathbb{E}[\log |Z|]\} + c_0 \geq \log^+ |\beta| + c'_0$. Finally, since both $h(\beta e^{j\Theta} + Z)$ and $\mathbb{E}[\log(|\beta + Z|)]$ are "close" to $\log^+ |\beta|$, they are "close" to each other due to the triangle inequality. This completes the proof of (9). ■

Lemma 8. For any $p, X > 0$, we have

$$|\log^+(pX) - \log^+ X| \leq |\log p|, \quad \text{and} \quad (12)$$

$$\mathbb{E}[\log^+ X] \leq p^{-1} \log^+(\mathbb{E}[X^p]) + p^{-1}. \quad (13)$$

Proof: To show (12), it is enough to verify that $\log^+(pX) - \log^+ X \leq \log p$ when $p \geq 1$ and $\log^+ X - \log^+(pX) \leq -\log p$ when $p < 1$, which completes the proof. The inequality (13) is based on Jensen's inequality. Specifically, we have $\mathbb{E}[\log^+ X] = p^{-1} \mathbb{E}[\log^+ X^p] \leq p^{-1} \mathbb{E}[\log(1 + X^p)] \leq p^{-1} \log(1 + \mathbb{E}[X^p]) \leq p^{-1} \log^+(\mathbb{E}[X^p]) + p^{-1}$. ■

IV. CAPACITY UPPER BOUND FOR MODEL A

The capacity $C(P)$ in (1) of a stationary and ergodic channel is upper-bounded by the capacity of the corresponding memoryless channel up to a constant term. Following the footsteps of [3], [5], we have

$$\begin{aligned} \frac{1}{N} I(\mathbf{X}^N; \mathbf{Y}^N) &= \frac{1}{N} \sum_{k=1}^N I(\mathbf{X}^N; \mathbf{Y}_k | \mathbf{Y}^{k-1}) \\ &\leq \frac{1}{N} \sum_{k=1}^N I(\mathbf{X}_k; \mathbf{Y}_k) + I(\Theta_N; \Theta^{N-1}) \\ &\leq \sup I(\mathbf{X}; \mathbf{Y}) + c_0, \end{aligned}$$

where $\sup I(\mathbf{X}; \mathbf{Y})$ is the capacity of a memoryless phase noise channel with the same temporal marginal distribution as the original channel, and the supremum is over all input distributions such that $\mathbb{E}[\|\mathbf{X}\|^2] \leq P$; using the fact that $I(\Theta_N; \Theta^{N-1}) = h(\Theta_N) - h(\Theta_N | \Theta^{N-1}) \leq \log(2\pi) - r_\Theta$ where r_Θ is the differential entropy rate of the phase noise process, we can set $c_0 = \log(2\pi) - r_\Theta$. Since we are mainly interested in the multiplexing gain, the constant c_0 does not matter, and it is thus without loss of optimality to consider the memoryless case in this section.

The main ingredients of the proof are the genie-aided bound and the duality upper bound. In the following, we detail the five steps that lead to Theorem 1.

A. Step 1: Genie-aided bound

Let us define the auxiliary random variable U as the index of the strongest input entry, i.e.,³

$$U \triangleq \arg \max_{1 \leq i \leq n_t} |X_i|.$$

Thus, we use X_U to denote the element in \mathbf{X} with the largest magnitude. It is obvious that $U \leftrightarrow \mathbf{X} \leftrightarrow \mathbf{Y}$ form a Markov chain, and that U does not contain more than $\log n_t$ bits. Assuming that a genie provides U to the receiver, we obtain the following upper bound

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &\leq I(\mathbf{X}; \mathbf{Y}, U) \\ &= I(\mathbf{X}; \mathbf{Y} | U) + I(U; \mathbf{X}) \\ &\leq I(\mathbf{X}; \mathbf{Y} | U) + H(U) \\ &\leq I(\mathbf{X}; \mathbf{Y} | U) + \log n_t. \end{aligned} \quad (14)$$

³When there are more than one such elements, we pick an arbitrary one.

B. Step 2: Canonical form

Definition 2 (Canonical channel). We define the canonical form u , $u = 1, \dots, n_t$, of the channel \mathbf{H} as

$$\mathbf{G}_{(u)} \triangleq \underbrace{\text{diag}(h_{1,u}^{-1}, \dots, h_{n_t,u}^{-1})}_{\mathbf{A}_u} \mathbf{H}. \quad (15)$$

Note that the elements in the u th column of $\mathbf{G}_{(u)}$ has normalized magnitudes. Now, with the information U from the genie, the receiver can convert the original channel into one of the canonical forms, namely, the form U .

$$\begin{aligned} I(\mathbf{X}; \mathbf{H}\mathbf{X} + \mathbf{Z} | U) &= I(\mathbf{X}; \mathbf{A}_U \mathbf{H}\mathbf{X} + \mathbf{A}_U \mathbf{Z} | U) \\ &\leq I(\mathbf{X}; \mathbf{A}_U \mathbf{H}\mathbf{X} + a\mathbf{Z} | U) \quad (16) \\ &= I(a^{-1}\mathbf{X}; a^{-1}\mathbf{G}_{(U)}\mathbf{X} + \mathbf{Z} | U) \\ &= I(\tilde{\mathbf{X}}; \mathbf{G}_{(U)}\tilde{\mathbf{X}} + \mathbf{Z} | U), \quad (17) \end{aligned}$$

where $a \triangleq \min_{k,u} |h_{k,u}^{-1}|$; (16) is due to the fact that reducing the additive noise increases the mutual information; we define

$$\tilde{\mathbf{X}} \triangleq a^{-1}\mathbf{X},$$

and accordingly,

$$\mathbf{W} \triangleq \mathbf{G}_{(u)}\tilde{\mathbf{X}} + \mathbf{Z}.$$

In the following, we focus on upper-bounding the mutual information $I(\tilde{\mathbf{X}}; \mathbf{W} | U)$. Note that

$$\begin{aligned} I(\tilde{\mathbf{X}}; \mathbf{W} | U) &= h(\mathbf{W} | U) - h(\mathbf{W} | \tilde{\mathbf{X}}, U) \\ &= h(\mathbf{W} | U) - h(\mathbf{W} | \tilde{\mathbf{X}}), \end{aligned} \quad (18)$$

where the last equality comes from the fact that U is a function of \mathbf{X} and thus a function of $\tilde{\mathbf{X}}$, since $\tilde{\mathbf{X}}$ is simply a scaled version of \mathbf{X} . Therefore, it is enough to lower-bound $h(\mathbf{W} | \tilde{\mathbf{X}})$ and upper-bound $h(\mathbf{W} | U)$ separately.

C. Step 3: Lower bound on $h(\mathbf{W} | \tilde{\mathbf{X}})$

Lemma 9. For model A, we have

$$h(\mathbf{W} | \tilde{\mathbf{X}}) \geq n_r \mathbb{E}[\log^+ |\tilde{X}_U|] + n_r \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H, \quad (19)$$

where \tilde{X}_U and \tilde{X}_V have the largest and second largest magnitudes in $\tilde{\mathbf{X}}$, respectively.

Proof: See Appendix B. ■

It is worth mentioning that the above bound depends not only on the strongest but also on the second strongest input of the channel.

D. Step 4: Upper bound on $h(\mathbf{W} | U)$

Upper-bounding $h(\mathbf{W} | U)$ by a non-trivial but tractable function of the input distribution is hard in general. A viable way for that purpose is through an auxiliary distribution, also called the duality approach. The duality upper bound was first proposed in [15] and [16] for discrete channels and then derived

for arbitrary channels in [3]. Namely, for any⁴ pdf $q(\mathbf{w})$, we have

$$\begin{aligned} h(\mathbf{W} | U) &= \mathbb{E}[-\log p(\mathbf{W} | U)] \\ &= \mathbb{E}[-\log q(\mathbf{W})] - \mathbb{E}_U [D(p_{\mathbf{W}|U=u} \| q)] \\ &\leq \mathbb{E}[-\log q(\mathbf{W})] \end{aligned} \quad (20)$$

due to the non-negativity of the Kullback-Leibler divergence $D(p_{\mathbf{W}|U=u} \| q)$. Hence, the key is to choose a proper auxiliary pdf $q(\mathbf{w})$ in order to obtain a tight upper bound on the capacity of our channel. The commonly used auxiliary distributions for MIMO channels are mostly related to the class of isotropic distributions [3], [5], [9]. Unfortunately, the isotropic distributions are not suitable in our case. To see this, let us assume that an isotropic output \mathbf{W} was indeed close to optimal. On the one hand, the pdf of an isotropic output \mathbf{W} would only depend on the norm $\|\mathbf{W}\|$ which would be dominated by the largest input entry X_U at high SNR. Therefore, the value of $\mathbb{E}[-\log q(\mathbf{W})]$ would be insensitive to the number of active input entries. On the other hand, the lower bound on the conditional entropy $h(\mathbf{W} | \tilde{\mathbf{X}})$ is increasing with *both* of the largest input entries X_U and X_V , according to (19). Therefore, with an isotropic distribution $q(\mathbf{w})$, the capacity upper bound $\mathbb{E}[-\log q(\mathbf{W})] - h(\mathbf{W} | \tilde{\mathbf{X}})$ would become larger when the second strongest input went to zero, i.e., only one transmit antenna was active. But this is in contradiction with the isotropic assumption, since if only one transmit antenna was active, then the output entries would be highly correlated and the output distribution would be *far* from being isotropic.

In light of the above discussion, we are led to think that a *good* choice of $q(\mathbf{w})$ should reflect not only the strongest input entry, but also the weaker ones. We adopt the following pdf built from the multivariate Gamma distribution in Definition 1,

$$\begin{aligned} q(\mathbf{w}) &= \frac{g\boldsymbol{\alpha}}{n_r!} |\hat{w}_1|^{2(\alpha_1-1)} \prod_{i=2}^{n_r} (|\hat{w}_i|^2 - |\hat{w}_{i-1}|^2)^{\alpha_i-1} \\ &\quad \cdot \exp(-\mu |\hat{w}_{n_r}|^2) \mu^{\alpha_1+\dots+\alpha_{n_r}}, \quad \mathbf{w} \in \mathbb{C}^{n_r}, \end{aligned} \quad (21)$$

where $\hat{w}_1, \dots, \hat{w}_{n_r}$ are the ordered version of w_i 's with increasing magnitudes. Essentially, we let each W_i be circularly symmetric and let the ordered version of $(|W_1|^2, \dots, |W_{n_r}|^2)$ follow the multivariate Gamma distribution defined in Definition 1. Applying (3) in Lemma 3 and the order statistics (whence the term $n_r!$) [20], we can obtain the pdf of \mathbf{W} as written in (21). Remarkably, the differences between $|W_i|^2$ and $|W_j|^2$, $i \neq j$, are introduced into the upper bound, which is crucial for bringing in the impact of individual input entries \tilde{X}_i 's other than the strongest entry as will be shown in the following.

Lemma 10. *By choosing $0 < \alpha_i < 1$, $i = 1, \dots, n_r$, and*

⁴Formally, we should state that the probability measure Q corresponding to the density $q(\mathbf{w})$ is such that $P(\cdot | U = u)$ is absolutely continuous with respect to Q . Throughout the paper, for brevity, we implicitly make the assumption to avoid such formalities.

$\mu = \min\{P^{-1}, 1\}$, we have for model A,

$$\begin{aligned} &\mathbb{E}[-\log q(\mathbf{W})] \\ &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \left((1 - \alpha_1) + \sum_{i=1}^{n_r} (1 - \alpha_i) \right) \mathbb{E}[\log^+ |\tilde{X}_U|] \\ &\quad + \sum_{i=2}^{n_r} (1 - \alpha_i) \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H, \end{aligned} \quad (22)$$

where \tilde{X}_U and \tilde{X}_V are the strongest and second strongest elements in $\tilde{\mathbf{X}}$, respectively.

Proof: The calculation is straightforward from the pdf (21), details are provided in Appendix C. ■

E. Step 5: Upper bound for model A

Combining (18), (19), (20), and (22) from the previous steps, we have

$$\begin{aligned} &I(\tilde{\mathbf{X}}; \mathbf{W} | U) \\ &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \left(1 - 2\alpha_1 - \sum_{i=2}^{n_r} \alpha_i \right) \mathbb{E}[\log^+ |\tilde{X}_U|] \\ &\quad + \left(\sum_{i=2}^{n_r} (1 - \alpha_i) - n_r \right) \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H \end{aligned} \quad (23)$$

$$\begin{aligned} &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \left(1 - 2\alpha_1 - \sum_{i=2}^{n_r} \alpha_i \right) \mathbb{E}[\log^+ |\tilde{X}_U|] + c_H \end{aligned} \quad (24)$$

$$\leq \left(\sum_{i=1}^{n_r} \alpha_i + \frac{1}{2} \right) \log^+ P + c'_H,$$

where the inequality (24) comes from removing the negative term in (23); to obtain the last inequality, we apply (13) in Lemma 8 with $p = 2$ and the power constraint $\mathbb{E}[|\tilde{X}_U|^2] \leq a^2 \mathbb{E}[\|\mathbf{X}\|^2] \leq a^2 P$.

Finally, we conclude from (14) and (17) that, for model A,

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &\leq I(\mathbf{X}; \mathbf{Y} | U) + c_0 \\ &\leq I(\tilde{\mathbf{X}}; \mathbf{W} | U) + c_0 \\ &\leq \left(\sum_{i=1}^{n_r} \alpha_i + \frac{1}{2} \right) \log^+ P + c_H \end{aligned}$$

which implies that the multiplexing gain is upper-bounded by

$$r_A \leq \sum_{i=1}^{n_r} \alpha_i + \frac{1}{2}, \quad \forall \boldsymbol{\alpha} \in (0, 1)^{n_r}.$$

By taking the infimum over $\boldsymbol{\alpha}$, we have $r_A \leq \frac{1}{2}$.

V. CAPACITY UPPER BOUND FOR MODEL B

In this section, we derive upper bounds for the three cases of model B, where the phase noises are on the transmitter and receiver sides of the channel. As in the previous section, it is enough to consider the memoryless case for our purpose.

A. Case B1: Transmit and receive phase noises

Note that the multiplexing gain of this case is upper-bounded by that of case B2 and case B3, since we can enhance the channel by providing the information on the transmit or receive phase noises to both the transmit and receiver. In other words, the upper bound $\min\{\frac{n_r}{2}, \frac{n_t}{2}, n_t - \frac{1}{2}\} = \min\{\frac{n_r}{2}, \frac{n_t}{2}\}$ is still valid for this case. In the following, we show that we can tighten the upper bound $\frac{n_r}{2}$ to $\frac{(n_r-2)^++1}{2}$ with the duality upper bound using the multi-variate Gamma distribution. The proof is in the same vein as the proof for model A. Specifically, the first four steps are exactly the same as for model A, except for Step 3 in which the conditional entropy has a different lower bound, as shown below.

Lemma 11. *For model B1, we have*

$$h(\mathbf{W} | \tilde{\mathbf{X}}) \geq n_r \mathbb{E}[\log^+ |\tilde{X}_U|] + \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H, \quad (25)$$

where \tilde{X}_U and \tilde{X}_V have the largest and second largest magnitudes in $\tilde{\mathbf{X}}$, respectively.

Proof: See Appendix B. ■

Applying (18), (20), (22), and (25), we have

$$\begin{aligned} I(\tilde{\mathbf{X}}; \mathbf{W} | U) &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \left(1 - 2\alpha_1 - \sum_{i=2}^{n_r} \alpha_i\right) \mathbb{E}[\log^+ |\tilde{X}_U|] \\ &\quad + \left(\sum_{i=2}^{n_r} (1 - \alpha_i) - 1\right) \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H \\ &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \mathbb{E}[\log^+ |\tilde{X}_U|] \\ &\quad + (n_r - 2)^+ \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H \\ &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + ((n_r - 2)^+ + 1) \mathbb{E}[\log^+ |\tilde{X}_U|] + c_H \\ &\leq \left(\frac{(n_r - 2)^+ + 1}{2} + \sum_{i=1}^{n_r} \alpha_i\right) \log^+ P + c'_H. \end{aligned} \quad (26)$$

Therefore, we conclude from (14), (17), and (26) that, for model B1,

$$I(\mathbf{X}; \mathbf{Y}) \leq \left(\sum_{i=1}^{n_r} \alpha_i + \frac{(n_r - 2)^+ + 1}{2}\right) \log^+ P + c'_H$$

which implies that the multiplexing gain is upper-bounded by

$$r_B \leq \sum_{i=1}^{n_r} \alpha_i + \frac{(n_r - 2)^+ + 1}{2}, \quad \forall \boldsymbol{\alpha} \in (0, 1)^{n_r}.$$

Taking the infimum over $\boldsymbol{\alpha}$, we have $r_B \leq \frac{(n_r-2)^++1}{2}$.

B. Case B2: Transmit phase noise

In this case, the received signal is $\mathbf{Y} = \mathbf{H}(e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X}) + \mathbf{Z}$. The channel is characterized by the random matrix $\mathbf{H} = \mathbf{H} \text{diag}\{e^{j\boldsymbol{\Theta}_T}\}$. We shall show that the upper bound is $\min\{\frac{n_t}{2}, \frac{n_r}{2}\}$. First, with more receive antennas than transmit

antennas, i.e., when $n_r \geq n_t$, we can inverse the channel without losing information,

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &= I(\mathbf{X}; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \mathbf{H}^\dagger \mathbf{Z}) \\ &\leq I(\mathbf{X}; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}}), \end{aligned} \quad (27)$$

where $\tilde{\mathbf{Z}} \sim \mathcal{CN}(0, \sigma_{\min}^2(\mathbf{H}^\dagger) \mathbf{I}_{n_t})$, with $\sigma_{\min}(\mathbf{H}^\dagger) > 0$ being the minimum singular value of \mathbf{H}^\dagger . Note that (27) is maximized when \mathbf{X} is circularly symmetric with n_t independent phases. To see this, we introduce a vector of independent and identically distributed (i.i.d.) phases $\boldsymbol{\Phi}$ uniformly distributed in $[0, 2\pi)^{n_t}$ and show that, for any \mathbf{X} ,

$$\begin{aligned} I(e^{j\boldsymbol{\Phi}} \circ \mathbf{X}; e^{j(\boldsymbol{\Theta}_T + \boldsymbol{\Phi})} \circ \mathbf{X} + \tilde{\mathbf{Z}}) &= h(e^{j(\boldsymbol{\Theta}_T + \boldsymbol{\Phi})} \circ \mathbf{X} + \tilde{\mathbf{Z}}) - h(e^{j(\boldsymbol{\Theta}_T + \boldsymbol{\Phi})} \circ \mathbf{X} + \tilde{\mathbf{Z}} | e^{j\boldsymbol{\Phi}} \circ \mathbf{X}) \\ &\geq h(e^{j(\boldsymbol{\Theta}_T + \boldsymbol{\Phi})} \circ \mathbf{X} + \tilde{\mathbf{Z}} | \boldsymbol{\Phi}) - h(e^{j(\boldsymbol{\Theta}_T + \boldsymbol{\Phi})} \circ \mathbf{X} + \tilde{\mathbf{Z}} | \mathbf{X}, \boldsymbol{\Phi}) \\ &= h(e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}}') - h(e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}}' | \mathbf{X}) \\ &= I(\mathbf{X}; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}}), \end{aligned}$$

where we use the fact that $\tilde{\mathbf{Z}}$ is circularly symmetric, and has the same distribution as $\tilde{\mathbf{Z}}' \triangleq e^{-j\boldsymbol{\Phi}} \tilde{\mathbf{Z}}$. Therefore, to derive an upper bound, it is without loss of optimality to assume that \mathbf{X} is circularly symmetric with m independent phases. With this assumption, we have

$$\begin{aligned} I(\mathbf{X}; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}}) &= I(|\mathbf{X}|; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}}) + I(\angle \mathbf{X}; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}} | |\mathbf{X}|) \\ &\leq I(|\mathbf{X}|; e^{j(\boldsymbol{\Theta}_T + \angle \mathbf{x})} \circ |\mathbf{X}| + \tilde{\mathbf{Z}}) + I(\angle \mathbf{X}; e^{j\boldsymbol{\Theta}_T} \circ \mathbf{X} | |\mathbf{X}|) \\ &\leq I(|\mathbf{X}|; e^{j(\boldsymbol{\Theta}_T + \angle \mathbf{x})} \circ |\mathbf{X}| + \tilde{\mathbf{Z}} | \boldsymbol{\Theta}_T + \angle \mathbf{x}) \\ &\quad + I(\angle \mathbf{X}; (\boldsymbol{\Theta}_T + \angle \mathbf{x})_{2\pi}) \\ &= I(|\mathbf{X}|; |\mathbf{X}| + \text{Re}\{\tilde{\mathbf{Z}}''\}) + h((\boldsymbol{\Theta}_T + \angle \mathbf{x})_{2\pi}) - h(\boldsymbol{\Theta}_T) \\ &\leq \frac{n_t}{2} \log^+ P + c_H + \log 2\pi - h(\boldsymbol{\Theta}_T), \end{aligned} \quad (28)$$

where the second inequality is obtain by providing $\boldsymbol{\Theta}_T + \angle \mathbf{x}$ to the output and the independence between $\boldsymbol{\Theta}_T + \angle \mathbf{x}$ and $|\mathbf{X}|$; the last inequality is from the capacity upper bound for a real-value Gaussian channel, and the fact that $(\boldsymbol{\Theta}_T + \angle \mathbf{x})_{2\pi}$ is uniformly distributed in $[0, 2\pi)$; we define $\tilde{\mathbf{Z}}'' \triangleq e^{-j(\boldsymbol{\Theta}_T + \angle \mathbf{x})} \circ \tilde{\mathbf{Z}}$. From (28), we get the upper bound $\frac{n_t}{2}$ of the pre-log.

In the following, we assume $n_r \leq n_t$, and follow closely to the proof for model A in Section IV-A. We first apply a genie-aided bound, by providing the set of indices of the n_r strongest inputs to the receiver. This information, also denoted by U , does not take more than $\log \binom{n_t}{n_r}$ bits. Then we also associate with each U a canonical form $\mathbf{G}_{(U)} = \mathbf{H}_U^{-1} \mathbf{H}$ where \mathbf{H}_U is the submatrix of \mathbf{H} with the columns corresponding to the n_r strongest entries, while $\mathbf{H}_{\bar{U}}$ corresponds to the rest of the columns. It follows that $\mathbf{H}_U^{-1} \mathbf{Y} = \mathbf{G}_{(U)} \mathbf{X} + \mathbf{H}_U^{-1} \mathbf{Z}$, and

$$\begin{aligned} I(\mathbf{X}; \mathbf{H}\mathbf{X} + \mathbf{Z} | U) &= I(\mathbf{X}; \mathbf{G}_{(U)} \mathbf{X} + \mathbf{H}_U^{-1} \mathbf{Z} | U) \\ &\leq I(\mathbf{X}; \mathbf{G}_{(U)} \mathbf{X} + a\mathbf{Z} | U) \\ &= I(a^{-1} \mathbf{X}; a^{-1} \mathbf{G}_{(U)} \mathbf{X} + \mathbf{Z} | U) \\ &= I(\tilde{\mathbf{X}}; \mathbf{G}_{(U)} \tilde{\mathbf{X}} + \mathbf{Z} | U), \end{aligned}$$

where $a \triangleq (\sigma_{\max}(\mathbf{H}))^{-1}$; we define $\tilde{\mathbf{X}} \triangleq a^{-1}\mathbf{X}$ and accordingly,

$$\mathbf{W} \triangleq \mathbf{G}_{(U)}\tilde{\mathbf{X}} + \mathbf{Z} = e^{j\Theta_{\mathbf{r},U}} \circ \tilde{\mathbf{X}}_U + \mathbf{H}_U^{-1}\mathbf{H}_{\bar{U}}(e^{j\Theta_{\mathbf{r},\bar{U}}} \circ \tilde{\mathbf{X}}_{\bar{U}}) + \mathbf{Z}.$$

The next step is to derive a lower bound on $h(\mathbf{W} | \tilde{\mathbf{X}})$,

$$\begin{aligned} h(\mathbf{W} | \tilde{\mathbf{X}}) &\geq h(\mathbf{W} | \tilde{\mathbf{X}}, \Theta_{\mathbf{T},\bar{U}}) \\ &= h(e^{j\Theta_{\mathbf{r},U}} \circ \tilde{\mathbf{X}}_U + \mathbf{Z} | \tilde{\mathbf{X}}, \Theta_{\mathbf{T},\bar{U}}) \\ &\geq \sum_{k=1}^{n_r} h(e^{j\Theta_{\mathbf{r},k}} \tilde{X}_k + Z_k | \tilde{\mathbf{X}}, \Theta_{\mathbf{T},\bar{U}}, \Theta_{\mathbf{T}}^{k-1}, Z^{k-1}) \\ &\geq \sum_{k=1}^{n_r} \mathbb{E}[\log^+ |\tilde{X}_k|] + c_0, \end{aligned} \quad (29)$$

where we assume that $U = \{1, \dots, n_r\}$ for notational convenience, and the last inequality is from Lemma 7.

Finally, we derive an upper bound on $h(\mathbf{W} | U)$ via duality using the following auxiliary distribution on the output \mathbf{W} ,

$$q(\mathbf{w}) = g_{\alpha} \mu^{\alpha_1 + \dots + \alpha_{n_r}} e^{-\mu \|\mathbf{w}\|^2} \prod_{i=1}^{n_r} |w_i|^{2(\alpha_i - 1)}, \quad (30)$$

where g_{α} is the normalization factor which only depends on α and n_r . Essentially, we let each W_i be independent and circularly symmetric with the squared magnitude following a single-variate Gamma distribution with parameter (μ, α_i) , as defined in (2) from Definition 1.

Lemma 12. *By choosing $0 < \alpha_i < 1$ and $\mu = \min\{P^{-1}, 1\}$, we have*

$$\begin{aligned} \mathbb{E}[-\log q(\mathbf{W})] &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \sum_{i=1}^{n_r} 2(1 - \alpha_i) \mathbb{E}[\log^+ |\tilde{X}_i|] + c_H, \end{aligned} \quad (31)$$

where we assume that $|\tilde{X}_1| \geq \dots \geq |\tilde{X}_{n_r}|$ for notational convenience.

Proof: The following is straightforward from (30),

$$\begin{aligned} \mathbb{E}[-\log q(\mathbf{W})] &\leq \sum_{i=1}^{n_r} \alpha_i \log \mu^{-1} + \mu \mathbb{E}[\|\mathbf{W}\|^2] \log e \\ &\quad + \sum_{i=1}^{n_r} 2(1 - \alpha_i) \mathbb{E}[\log |W_i|] + c_0 \\ &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \sum_{i=1}^{n_r} (1 - \alpha_i) \mathbb{E}[\log |W_i|^2] + c'_H. \end{aligned} \quad (32)$$

Note that $|W_i|^2 \leq 2|\tilde{X}_i|^2 + 2\sigma_{\max}^2(\mathbf{H}_U^{-1}\mathbf{H}_{\bar{U}})\|\tilde{\mathbf{X}}_{\bar{U}}\|^2 + 2|Z_i|^2$. From the definition of U , we have $\|\tilde{\mathbf{X}}_{\bar{U}}\|^2 \leq (n_t - n_r)|\tilde{X}_i|^2$, $\forall i \in U$. Thus, $|W_i|^2 \leq (2 + |c''_H|)|\tilde{X}_i|^2 + 2|Z_i|^2$. Applying Jensen's inequality on the expectation over Z_i , we get $\mathbb{E}[\log |W_i|^2] \leq \mathbb{E}[\log((2 + |c''_H|)|\tilde{X}_i|^2 + 2)] \leq 2\mathbb{E}[\log^+ |\tilde{X}_i|] + c''_H$. Plugging it back to (32), we readily have (31). ■

Finally, putting together (29) and (31), we obtain

$$\begin{aligned} I(\tilde{\mathbf{X}}; \mathbf{W} | U) &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \sum_{i=1}^{n_r} (1 - 2\alpha_i) \mathbb{E}[\log^+ |\tilde{X}_i|] + c_H \\ &\leq \sum_{i=1}^{n_r} \alpha_i \log^+ P + \sum_{i=1}^{n_r} (1 - 2\alpha_i)^+ \mathbb{E}[\log^+ |\tilde{X}_i|] + c_H \\ &\leq \frac{1}{2} \sum_{i=1}^{n_r} (2\alpha_i + (1 - 2\alpha_i)^+) \log^+ P + c'_H, \end{aligned}$$

where, to obtain the last inequality, we apply (13) in Lemma 8 with $p = 2$, and the power constraint $\mathbb{E}[|\tilde{X}_i|^2] \leq a^2 \mathbb{E}[\|\mathbf{X}\|^2] \leq a^2 P$. Therefore, the multiplexing gain is upper-bounded by

$$\frac{1}{2} \sum_{i=1}^{n_r} (2\alpha_i + (1 - 2\alpha_i)^+), \quad \forall \alpha \in (0, 1)^{n_r}.$$

Taking the infimum over α , we get $\frac{n_r}{2}$.

C. Case B3: Receive phase noise

First it is not hard to show the upper bound $n_r - \frac{1}{2}$. It is enough to provide the $n_r - 1$ relative angles, $\{\Theta_{\mathbf{R},k} - \Theta_{\mathbf{R},1}\}_{k=2 \dots n_r}$, to the receiver. The channel is now equivalent to the case with common phase noise $\Theta_{\mathbf{R},1}$. Then we can apply Proposition 1, since

$$\begin{aligned} h(\Theta_{\mathbf{R},1} | \{\Theta_{\mathbf{R},k} - \Theta_{\mathbf{R},1}\}_{k=2 \dots n_r}) &= h(\Theta_{\mathbf{R}}) - h(\{\Theta_{\mathbf{R},k} - \Theta_{\mathbf{R},1}\}_{k=2 \dots n_r}) \\ &\geq h(\Theta_{\mathbf{R}}) - (n_r - 1) \log(2\pi) \\ &\geq -\infty. \end{aligned}$$

Next, we show the upper bound $\frac{n_r}{2}$. To that end, we write

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &= I(\mathbf{H}\mathbf{X}; |\mathbf{Y}|) + I(\mathbf{H}\mathbf{X}; \angle \mathbf{Y} | |\mathbf{Y}|) \\ &= I(\mathbf{H}\mathbf{X}; |\mathbf{Y}|) + I(\mathbf{H}\mathbf{X}; (\angle \mathbf{H}\mathbf{X} + \Theta_{\mathbf{R}})_{2\pi} | |\mathbf{Y}|) \\ &= I(\mathbf{H}\mathbf{X}; |\mathbf{Y}|) + h((\angle \mathbf{H}\mathbf{X} + \Theta_{\mathbf{R}})_{2\pi} | |\mathbf{Y}|) \\ &\quad - h((\angle \mathbf{H}\mathbf{X} + \Theta_{\mathbf{R}})_{2\pi} | |\mathbf{Y}|, \mathbf{H}\mathbf{X}) \\ &\leq I(\mathbf{H}\mathbf{X}; |\mathbf{Y}|) + n_r \log(2\pi) \\ &\quad - h((\angle \mathbf{H}\mathbf{X} + \Theta_{\mathbf{R}})_{2\pi} | |\mathbf{Y}|, \mathbf{H}\mathbf{X}, \angle \mathbf{H}\mathbf{X}) \\ &= I(\mathbf{H}\mathbf{X}; |\mathbf{Y}|) + n_r \log(2\pi) - h(\Theta_{\mathbf{R}}), \end{aligned}$$

where we define $\mathbf{Z}' \triangleq e^{-j\Theta_{\mathbf{R}}} \circ \mathbf{Z}$ which is independent of $\Theta_{\mathbf{R}}$ since \mathbf{Z} is circularly symmetric; the last equality follows since $\mathbf{Y} = e^{j\Theta_{\mathbf{R}}} \circ (\mathbf{H}\mathbf{X} + \mathbf{Z}')$ and thus $\Theta_{\mathbf{R}}$ is independent of $(|\mathbf{Y}|, \mathbf{H}\mathbf{X} + \mathbf{Z}', \mathbf{H}\mathbf{X})$. It remains to show that $I(\mathbf{H}\mathbf{X}; |\mathbf{Y}|) \leq \frac{n_r}{2} \log^+ P + c_H$. To prove this, it is enough to apply $h(|\mathbf{Y}|) \leq \frac{n_r}{2} \log^+ P + c_H$ and to use the fact that $h(|\mathbf{Y}| | \mathbf{H}\mathbf{X}) = \sum_{k=1}^{n_r} h(|Y_k| | \mathbf{H}\mathbf{X})$ is lower-bounded by some constant according to (9) in Lemma 7.

VI. CAPACITY LOWER BOUND FOR MODEL B

In this section, we derive a lower bound on the capacity of model B. For simplicity, we consider the class of memoryless Gaussian input distributions. Although the optimal input distribution has been proved to be discrete in [6], the use of a simple Gaussian input provides tight lower bounds on the pre-log, which is enough for our purpose here. In the following, we only consider the memoryless phase noise channel which can be shown to have a lower capacity than the general stationary and ergodic channel when memoryless input is used. To see this, we write

$$\begin{aligned} I(\mathbf{X}^N; \mathbf{Y}^N) &= h(\mathbf{X}^N) - h(\mathbf{X}^N | \mathbf{Y}^N) \\ &= \sum_{t=1}^N h(\mathbf{X}_t) - \sum_{t=1}^N h(\mathbf{X}_t | \mathbf{X}^{t-1}, \mathbf{Y}^N) \\ &\geq \sum_{t=1}^N h(\mathbf{X}_t) - \sum_{t=1}^N h(\mathbf{X}_t | \mathbf{Y}_t) \\ &= \sum_{t=1}^N I(\mathbf{X}_t; \mathbf{Y}_t) \\ &= NI(\mathbf{X}; \mathbf{Y}). \end{aligned}$$

Thus, we focus on the single-letter mutual information $I(\mathbf{X}; \mathbf{Y})$ in the rest of the section. As in the previous section, we investigate the three cases separately.

A. Case B1: Transmit and receive phase noises

In this case, we use all the inputs with equal power, i.e., $\mathbf{X} \sim \mathcal{CN}(0, \frac{P}{n_t} \mathbf{I}_{n_t})$. For convenience, let us rewrite the received signal as

$$\begin{aligned} \mathbf{Y} &= e^{j\Theta_R} \circ (\mathbf{H}(e^{j\Theta_T} \circ \mathbf{X})) + \mathbf{Z} \\ &= \sqrt{\frac{P}{n_t}} e^{j\tilde{\Theta}_R} \circ (\mathbf{H}(e^{j\tilde{\Theta}_T} \circ \mathbf{X}_0)) + \mathbf{Z} \\ &= \sqrt{\frac{P}{n_t}} e^{j\tilde{\Theta}_R} \circ \tilde{\mathbf{Y}} + \mathbf{Z} = \sqrt{\frac{P}{n_t}} \tilde{\mathbf{Y}} + \mathbf{Z}, \end{aligned}$$

where $\mathbf{X}_0 \sim \mathcal{CN}(0, \mathbf{I}_{n_t})$ is the normalized version of \mathbf{X} ; $\tilde{\Theta}_R \triangleq \Theta_R + \Theta_{T,1}$ and $\tilde{\Theta}_T \triangleq \Theta_T - \Theta_{T,1}$. Note that $\tilde{\Theta}_{T,1} = 0$ by definition and $h(\tilde{\Theta}_R) > -\infty$. The mutual information of interest can be written as

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &= I(\mathbf{X}, \tilde{\Theta}_T; \mathbf{Y}) - I(\tilde{\Theta}_T; \mathbf{Y} | \mathbf{X}) \\ &= h(\mathbf{Y}) - h(\mathbf{Y} | \mathbf{X}, \tilde{\Theta}_T) - I(\tilde{\Theta}_T; \mathbf{Y} | \mathbf{X}). \end{aligned} \quad (33)$$

First the following lemma, which provides a lower bound on $h(\mathbf{Y})$ in (33), is crucial for the achievability proof.

Lemma 13. *With receive phase noise such that $h(\Theta_R) > -\infty$, we have*

$$h(\mathbf{Y}) \geq \left(\frac{n_r}{2} + \frac{1}{2} \min\{n_r, 2n_t - 1\} \right) \log^+ P + c_H. \quad (34)$$

Proof: See Appendix D. ■

Next, we derive upper bounds on the two negative terms in (33) as follows. The conditional differential entropy can be

upper-bounded as

$$\begin{aligned} h(\mathbf{Y} | \mathbf{X}, \tilde{\Theta}_T) &\leq \sum_{k=1}^{n_r} h(Y_k | \mathbf{X}, \tilde{\Theta}_T) \\ &\leq \sum_{k=1}^{n_r} \mathbb{E} \left[\log^+ \left| \sqrt{\frac{P}{n_t}} \mathbf{h}_k^T(e^{j\tilde{\Theta}_T} \circ \mathbf{X}_0) \right| \right] + c_0 \\ &\leq \frac{n_r}{2} \log^+ P + c_H, \end{aligned} \quad (35)$$

where the second inequality is due to Lemma 7 and the third inequality is from Lemma 8 and the power constraint $\mathbb{E} \left[\left| \mathbf{h}_k^T(e^{j\tilde{\Theta}_T} \circ \mathbf{X}_0) \right|^2 \right] = n_t^{-1} \|\mathbf{h}_k\|^2 P$. And

$$\begin{aligned} I(\tilde{\Theta}_T; \mathbf{Y} | \mathbf{X}) &\leq I(\tilde{\Theta}_T; \mathbf{Y}, \tilde{\Theta}_R | \mathbf{X}) \\ &= I(\tilde{\Theta}_T; \mathbf{Y} | \mathbf{X}, \tilde{\Theta}_R) + I(\tilde{\Theta}_T; \tilde{\Theta}_R) \\ &= I(\tilde{\Theta}_T; e^{j\tilde{\Theta}_T} \circ \mathbf{X} + \mathbf{H}^\dagger \mathbf{Z} | \mathbf{X}, \tilde{\Theta}_R) + c_0 \\ &\leq I(\tilde{\Theta}_T; e^{j\tilde{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}} | \mathbf{X}, \tilde{\Theta}_R) + c_0 \\ &= h(e^{j\tilde{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}} | \mathbf{X}, \tilde{\Theta}_R) \\ &\quad - h(e^{j\tilde{\Theta}_T} \circ \mathbf{X} + \tilde{\mathbf{Z}} | \mathbf{X}, \tilde{\Theta}_R, \tilde{\Theta}_T) + c_0 \\ &\leq \frac{n_t - 1}{2} \log^+ P + c_H, \end{aligned} \quad (38)$$

where $\tilde{\mathbf{Z}} \sim \mathcal{CN}(0, \sigma_{\min}^2(\mathbf{H}^\dagger) \mathbf{I}_{n_t})$, with $\sigma_{\min}(\mathbf{H}^\dagger)$ being the minimum singular value of \mathbf{H}^\dagger ; to obtain the last inequality, we use the fact that $\tilde{\Theta}_{T,1} = 0$ and apply Lemma 7 then Lemma 8 for the rest of the $n_t - 1$ entries of $\tilde{\Theta}_T$.

Plugging (36), (38), and (34) into (33), we obtain

$$I(\mathbf{X}; \mathbf{Y}) \geq \frac{1}{2} \min\{n_r - n_t + 1, n_t\} \log^+ P + c_H.$$

Note that the above lower bound holds when we substitute n_t by any $n'_t \leq n_t$, i.e., by activating only n'_t transmit antennas. It is clear that when $n_r - n_t + 1 \geq n_t$, i.e., $n_r \geq 2n_t - 1$, we should let $n'_t = n_t$. Otherwise, we should decrease n'_t to balance between $n_r - n'_t + 1$ and n'_t , which gives $n'_t = \lfloor \frac{n_r + 1}{2} \rfloor$. This completes the proof of the lower bound for model B1.

B. Case B2: Transmit phase noise

In this case, we use $n'_t \triangleq \min\{n_t, n_r\}$ input antennas and deactivate the remaining ones. The active inputs, denoted by \mathbf{X}' , are i.i.d. Gaussian, i.e., $\mathbf{X}' \sim \mathcal{CN}(0, \frac{P}{n'_t} \mathbf{I}_{n'_t})$. We rewrite the output vector as $\mathbf{Y} = \mathbf{H}'(e^{j\Theta_T} \circ \mathbf{X}') + \mathbf{Z}$ where $\mathbf{H}' \in \mathbb{C}^{n_r \times n'_t}$ is the submatrix of \mathbf{H} corresponding to the active inputs, and Θ_T' is similarly defined. It follows that $I(\mathbf{X}'; \mathbf{Y}) = I(\mathbf{X}'; (\mathbf{H}')^\dagger \mathbf{Y})$. Then we have $h((\mathbf{H}')^\dagger \mathbf{Y}) = h(e^{j\Theta_T'} \circ \mathbf{X}' + (\mathbf{H}')^\dagger \mathbf{Z}) = n'_t \log^+ P + c_H$ and $h((\mathbf{H}')^\dagger \mathbf{Y} | \mathbf{X}') = h(e^{j\Theta_T'} \circ \mathbf{X}' + (\mathbf{H}')^\dagger \mathbf{Z} | \mathbf{X}') \leq h(e^{j\Theta_T'} \circ \mathbf{X}' + \sigma_{\max}((\mathbf{H}')^\dagger) \mathbf{Z} | \mathbf{X}')$. The latter is further upper-bounded by $\sum_{k=1}^{n'_t} \mathbb{E}[\log^+ |X_k|] + c_H \leq \frac{n'_t}{2} \log^+ P + c'_H$ according to (8) in Lemma 7 and (13) in Lemma 8. This shows the lower bound $\frac{1}{2} \min\{n_t, n_r\}$ on the multiplexing gain.

C. Case B3: Receive phase noise

As in Case B1, we let $\mathbf{X} \sim \mathcal{CN}(0, \frac{P}{n_t} \mathbf{I}_{n_t})$. First $h(\mathbf{Y})$ is lower-bounded in Lemma 13. Next, it readily follows from

(35) to (36) that

$$h(\mathbf{Y} | \mathbf{X}) \leq \frac{n_r}{2} \log^+ P + c_H, \quad (39)$$

since we are in the same situation as in Case B1 when $\tilde{\Theta}_T$ is known. Finally, combining (34) and (39), we obtain a lower bound on the mutual information

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &= h(\mathbf{Y}) - h(\mathbf{Y} | \mathbf{X}) \\ &\geq \frac{1}{2} \min\{n_r, 2n_t - 1\} \log^+ P + c_H \end{aligned}$$

which provides the desired multiplexing gain.

VII. CONCLUSIONS AND DISCUSSIONS

In this work, we investigated the discrete-time stationary and ergodic $n_r \times n_t$ MIMO phase noise channel. We characterized the exact multiplexing gain when phase noises are on the individual paths and when phase noises are at either side of the channel. With both transmit and receive phase noises, upper and lower bounds have been derived. In particular, the upper bound results in this paper have been obtained via the duality using a newly introduced multi-variate Gamma distribution.

For model B1, the upper and lower bounds derived in this paper do not match for $n_r \in [4 : 2n_t - 2]$. We conjecture that the upper bound $\frac{1}{2} \min\{n_t, n_r - 1\}$ is indeed loose. Let us recall that the upper bound is obtained by lower-bounding $h(\mathbf{W} | \tilde{\mathbf{X}})$ with (25), and by upper-bounding $\mathbb{E}[-\log q(\mathbf{W})]$ with $q(\mathbf{w})$ being the multi-variate Gamma distribution. We believe that both bounds are loose for model B1 in general. First we can show that

$$\begin{aligned} h(\mathbf{W} | \tilde{\mathbf{X}}) &\leq n_r \mathbb{E}[\log^+ |\tilde{X}_U|] + \mathbb{E}[\log^+ |\tilde{X}_V|] \\ &\quad + \sum_{k \notin \{U, V\}} \mathbb{E}[\log^+ |\tilde{X}_k|] + c_H. \end{aligned} \quad (40)$$

To see this, we can first write $h(\mathbf{W} | \tilde{\mathbf{X}}) = h(\mathbf{W} | \tilde{\mathbf{X}}, \tilde{\Theta}_T) + I(\tilde{\Theta}_T; \mathbf{W} | \tilde{\mathbf{X}})$, then upper-bound the first term with $n_r \mathbb{E}[\log^+ |\tilde{X}_U|] + c'_H$ by following closely the steps as in (35)-(36), and upper-bound the second term with $\sum_{k \neq U} \mathbb{E}[\log^+ |\tilde{X}_k|] + c''_H$ by following closely the steps as in (37)-(38). As compared to the lower bound (25), the upper bound (40) differs only in the terms involving \tilde{X}_k , $k \notin \{U, V\}$. In the following, we argue that even if the lower bound $h(\mathbf{W} | \tilde{\mathbf{X}})$ was the RHS of (40) – which is the largest that one could get as lower bound since it is also an upper bound – we still would not be able to tighten the multiplexing gain upper bound $\frac{1}{2} \min\{n_r - 1\}$ with the same choice of auxiliary distribution $q(\mathbf{w})$. In other words, for the given $q(\mathbf{w})$, (25) is tight enough with respect to the upper bound on $\mathbb{E}[-\log q(\mathbf{W})] - h(\mathbf{W} | \tilde{\mathbf{X}})$. To prove this, it is enough to observe that $\mathbb{E}[-\log q(\mathbf{W})]$ does not involve any terms of $\tilde{\mathbf{X}}$ other than \tilde{X}_U and \tilde{X}_V in such a way to change the high SNR behavior, whereas $h(\mathbf{W} | \tilde{\mathbf{X}})$ is increasing with the strength of each \tilde{X}_k . Therefore, the maximization of $\mathbb{E}[-\log q(\mathbf{W})] - h(\mathbf{W} | \tilde{\mathbf{X}})$ over $\tilde{\mathbf{X}}$ will always put all \tilde{X}_k , $k \notin \{U, V\}$, to zero, even if $h(\mathbf{W} | \tilde{\mathbf{X}})$ hits the highest value (40). To sum up, if the current upper bound $\frac{1}{2} \min\{n_t, n_r - 1\}$ was indeed loose as we conjecture, one would have to first find

a new auxiliary distribution $q(\mathbf{w})$ in order to get a tighter upper bound. In particular, the new auxiliary distribution should be such that $\mathbb{E}[-\log q(\mathbf{W})]$ depends on \tilde{X}_k , $k \notin \{U, V\}$ at high SNR in a non-trivial way. With such a distribution, the second challenge is to find a lower bound on $h(\mathbf{W} | \tilde{\mathbf{X}})$ that also depends on \tilde{X}_k , $k \notin \{U, V\}$, in a non-trivial way. In fact, we conjecture that (40) holds with equality.

For model B2 and B3, the results have the following alternative chain rule interpretation. With transmit phase noise (model B2), the mutual information can be written as $I(\mathbf{X}; \mathbf{Y}) = I(\mathbf{X}, \Theta_T; \mathbf{Y}) - I(\Theta_T; \mathbf{Y} | \mathbf{X})$, where the first term scales as $\min\{n_t, n_r\} \log P$ as if the phase noise were part of the transmitted signal whereas the second part scales as $\frac{1}{2} \min\{n_t, n_r\} \log P$ as if Θ were the input with a fixed distribution and \mathbf{X} were the “fading” known at the receiver side. With receive phase noise (model B), the mutual information can be written differently as $I(\mathbf{X}; \mathbf{Y}) = I(\mathbf{X}; \mathbf{Y} | \Theta_R) - I(\Theta_R; \mathbf{X} | \mathbf{Y})$. Here the first term corresponds to the rate when the phase noise is known, while the second term can be considered as the rate of a “reverse” channel with input Θ_R , output \mathbf{X} , and known fading \mathbf{Y} . In both cases, the original problem of characterizing $I(\mathbf{X}; \mathbf{Y})$ boils down to subproblems involving channels without phase noise (i.e., $I(\mathbf{X}, \Theta_T)$ and $I(\mathbf{X}; \mathbf{Y} | \Theta_R)$) and communications with fixed phase signaling (i.e., $I(\Theta_T; \mathbf{Y} | \mathbf{X})$ and $I(\Theta_R; \mathbf{X} | \mathbf{Y})$).

There are a few interesting future directions. First, it is possible to extend the results to multi-user channels and study the impact of phase noise to such systems. Second, the lower bound on model B1 suggests the following dimension counting argument: one can recover n_t real information with $2n_t - 1$ real observations, since the remaining $n_t - 1$ dimensions are occupied by the $n_t - 1$ relative phase noises. How to design decoding algorithms that “solve” efficiently the $2n_t - 1$ non-linear equations is a question of both theoretical and practical importance. Finally, a more refined analysis should lead to tighter upper and lower bounds on the capacity, beyond the pre-log characterization.

APPENDIX

A. Proof of Proposition 1

With common phase noise, we can perform unitary precoding without losing information, and the channel is equivalent to a parallel channel with common phase noise $\mathbf{Y}_t = e^{j\Theta_t} \Sigma \mathbf{x}_t + \mathbf{Z}_t = e^{j\Theta_t} \sigma \circ \mathbf{x}_t + \mathbf{Z}_t$, where Σ is a diagonal matrix with the $\min\{n_t, n_r\}$ non-zero singular values of the matrix \mathbf{H} and σ is a vector of these elements. From [9], we know that the multiplexing gain of a $M \times M$ channel is upper-bounded by $M - \frac{1}{2}$. This upper bound applies here with $M = \min\{n_t, n_r\}$. The lower bound is achieved by using the Gaussian memoryless input $\mathbf{X}_t \sim \mathcal{CN}(0, \frac{P}{n_t} \mathbf{I}_{n_t})$, from which we have $I(\mathbf{X}^N; \mathbf{Y}^N) = h(\mathbf{Y}^N) - h(\mathbf{Y}^N | \mathbf{X}^N)$ with $h(\mathbf{Y}^N) = Nh(\mathbf{Y}) = N \min\{n_t, n_r\} \log^+ P + Nc_H$ and $h(\mathbf{Y}^N | \mathbf{X}^N) \leq Nh(\mathbf{Y} | \mathbf{X}) = Nh(e^{j\Theta} \sigma \circ \mathbf{X} + \mathbf{Z} | \mathbf{X})$. Applying a unitary transformation on $e^{j\Theta} \sigma \circ \mathbf{X} + \mathbf{Z}$, we obtain $Nh(e^{j\Theta} \sigma \circ \mathbf{X} + \mathbf{Z} | \mathbf{X}) = Nh(e^{j\Theta} \|\sigma \circ \mathbf{X}\| + \mathbf{Z}'_1 | \mathbf{X}) + N \sum_{k=2}^M h(\mathbf{Z}'_k) \leq N\mathbb{E}[\log^+ \|\sigma \circ \mathbf{X}\|] + Nc'_H \leq \frac{N}{2} \log^+ P + Nc'_H$ where \mathbf{Z}' is the rotated version of \mathbf{Z} and remains spatially white, the first inequality is from Lemma 7 and the second one is from

Lemma 8. Finally, we have $\frac{1}{N}I(\mathbf{X}^N; \mathbf{Y}^N) \geq (\min\{n_t, n_r\} - \frac{1}{2}) \log^+ P + c_H''$, which completes the proof.

B. Proof of Lemma 9 and 11

In the following we shall derive the lower bounds (19) and (25) on the conditional differential entropy $h(\mathbf{W} | \tilde{\mathbf{X}})$ for model A and model B1, respectively.

First we shall show that, for both models,

$$h(W_i | \tilde{\mathbf{X}}) \geq \mathbb{E}[\log^+ |\tilde{X}_U|] + \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H. \quad (41)$$

To that end, we analyze $h(W_i | \tilde{\mathbf{X}} = \mathbf{x})$ with $|x_1| > |x_2| > \dots > |x_{n_t}| \geq 0$ without loss of generality, i.e., we assume that $U = 1$ and $V = 2$. A lower bound of $h(W_i | \tilde{\mathbf{X}} = \mathbf{x})$ can be obtained by considering the following cases separately.

- When $|x_1| \leq 1$,

$$\begin{aligned} h(W_i | \tilde{\mathbf{X}} = \mathbf{x}) &\geq h(W_i | \tilde{\mathbf{X}} = \mathbf{x}, \Theta) \\ &= h(Z_i) \\ &= \log(\pi e). \end{aligned}$$

- When $|x_1| \geq 1$ and $|x_2| \leq 1$,

$$\begin{aligned} h(W_i | \tilde{\mathbf{X}} = \mathbf{x}) &\geq h(W_i | \tilde{\mathbf{X}} = \mathbf{x}, \Theta_{i,2}, \dots, \Theta_{i,n_t}) \\ &= h(g_{i1} e^{j\Theta_{i,1}} x_1 + Z_i | \Theta_{i,2}, \dots, \Theta_{i,n_t}) \\ &\geq \log^+ |g_{i1} x_1| + c_0 \\ &\geq \log^+ |x_1| + c_H, \end{aligned}$$

where $g_{i1} e^{j\Theta_{i,1}}$ is from the matrix $\mathbf{G}_{(1)}$ defined in (15) since $U = 1$ by assumption; the second inequality is from Lemma 7 and the third inequality is from Lemma 8.

- When $|x_1| \geq 1$ and $|x_2| \geq 1$,

$$\begin{aligned} h(W_i | \tilde{\mathbf{X}} = \mathbf{x}) &\geq h(W_i | \tilde{\mathbf{X}} = \mathbf{x}, \Theta_{i,3}, \dots, \Theta_{i,n_t}, Z_i) \\ &= h(e^{j\Theta_{i,1}} g_{i1} x_1 + e^{j\Theta_{i,2}} g_{i2} x_2 | \Theta_{i,3}, \dots, \Theta_{i,n_t}) \\ &\geq h(e^{j\Theta_{i,1}} g_{i1} x_1 + e^{j\Theta_{i,2}} g_{i2} x_2 | \Theta_{i,3}, \dots, \Theta_{i,n_t}, \Omega) \\ &= \mathbb{E} \log(|g_{i1} g_{i2} x_1 x_2 \sin(\Theta_{i,1} - \Theta_{i,2} + \phi)|) \\ &\quad + h(\Theta_{i,1}, \Theta_{i,2} | \Theta_{i,3}, \dots, \Theta_{i,n_t}, \Omega) \quad (42) \\ &\geq \log |x_1| + \log |x_2| \\ &\quad + \mathbb{E} \log(|\sin(\Theta_{i,1} - \Theta_{i,2} + \phi)|) + c_H \\ &\geq \log^+ |x_1| + \log^+ |x_2| + c_H', \end{aligned}$$

where the first inequality is from conditioning reduces entropy; we partition $[0, 2\pi)^2$ in such a way that $e^{j\Theta_{i,1}} g_{i1} x_1 + e^{j\Theta_{i,2}} g_{i2} x_2$ is a bijective function of $(\Theta_{i,1}, \Theta_{i,2})$ in each partition indexed by Ω which takes a finite number of values; then we applied the change of variables from Lemma 2 and obtain (42) with $\phi \triangleq \angle_{g_{i1} x_1} - \angle_{g_{i2} x_2}$; finally, we use the fact that $|g_{i1} g_{i2}|$ is bounded for almost every \mathbf{H} and the application of Lemma 5 to get the last inequality. Note that $\log |x_k| = \log^+ |x_k|$ for $k = 1, 2$ by assumption.

Combining the three cases above and taking expectation over $\tilde{\mathbf{X}}$, we get (41).

1) *Proof of the lower bound (19) for model A:* For model A, we have

$$\begin{aligned} h(\mathbf{W} | \tilde{\mathbf{X}}) &= \sum_{i=1}^{n_r} h(W_i | \tilde{\mathbf{X}}, W^{i-1}) \\ &\geq \sum_{i=1}^{n_r} h(W_i | \tilde{\mathbf{X}}, W^{i-1}, \{\Theta_{l,1}, \dots, \Theta_{l,n_t}\}_{l < i}) \\ &= \sum_{i=1}^{n_r} h(W_i | \tilde{\mathbf{X}}, \{\Theta_{l,1}, \dots, \Theta_{l,n_t}\}_{l < i}) \quad (43) \\ &\geq n_r \mathbb{E}[\log^+ |\tilde{X}_U|] + n_r \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H, \end{aligned}$$

where (43) is from the fact that W_i only depends on W^{i-1} through the input $\tilde{\mathbf{X}}$ and the phase noises $\{\Theta_{l,1}, \dots, \Theta_{l,n_t}\}_{l < i}$; where the last inequality is from a modified version of (41) by introducing $\{\Theta_{l,1}, \dots, \Theta_{l,n_t}\}_{l < i}$ in the condition.

2) *Proof of the lower bound (25) for model B1:* For model B1, we write

$$h(\mathbf{W} | \tilde{\mathbf{X}}) = h(W_1 | \tilde{\mathbf{X}}) + \sum_{i=2}^{n_r} h(W_i | \tilde{\mathbf{X}}, W^{i-1}), \quad (44)$$

where, according to (41), the first term is lower-bounded by

$$h(W_1 | \tilde{\mathbf{X}}) \geq \mathbb{E}[\log^+ |\tilde{X}_U|] + \mathbb{E}[\log^+ |\tilde{X}_V|] + c_H. \quad (45)$$

In the following, we derive a lower bound on the second term. Let $B_i \triangleq \sum_{k=1}^{n_t} g_{ik} \tilde{X}_k e^{j\Theta_{i,k}}$ where g_{ik} is the channel coefficient without phase noise from the canonical form U defined in (15). Then

$$\begin{aligned} h(W_i | \tilde{\mathbf{X}}, W^{i-1}) &= h(e^{j\Theta_{i,i}} B_i + Z_i | \tilde{\mathbf{X}}, W^{i-1}) \\ &\geq h(e^{j\Theta_{i,i}} B_i + Z_i | \tilde{\mathbf{X}}, W^{i-1}, B_i, \Theta_{\mathbf{T}}, \Theta_{\mathbf{R}}^{i-1}) \\ &= h(e^{j\Theta_{i,i}} B_i + Z_i | B_i, \Theta_{\mathbf{T}}, \Theta_{\mathbf{R}}^{i-1}) \\ &\geq \mathbb{E}[\log^+ |B_i|] + c_0 \quad (46) \\ &= \mathbb{E}_{\tilde{\mathbf{X}}} [\mathbb{E}[\log^+ |B_i| | \tilde{\mathbf{X}}]] + c_0, \quad (47) \end{aligned}$$

where the first inequality is from conditioning reduces entropy; (46) is from Lemma 7. The conditional expectation can be lower-bounded as follows

$$\begin{aligned} \mathbb{E}[\log |B_i| | \tilde{\mathbf{X}}] &= \frac{1}{2} \mathbb{E} \left[\log \left(\left| \sum_{k=1}^{n_t} g_{ik} \tilde{X}_k e^{j(\Theta_{i,k} + \tilde{\Phi}_{ik})} \right|^2 \right) \right] \\ &\geq \frac{1}{2} \inf_{\mathbf{x} \in \mathbb{R}^{n_t}: \|\mathbf{x}\|=1} \mathbb{E} \left[\log \left(\left| \sum_{k=1}^{n_t} |g_{ik}| e^{j(\Theta_{i,k} + \tilde{\Phi}_{ik})} x_k \right|^2 \right) \right] \\ &\quad + \log \|\tilde{\mathbf{X}}\| \\ &\geq \frac{1}{2} \inf_{\mathbf{x} \in \mathbb{R}^{n_t}: \|\mathbf{x}\|=1} \mathbb{E} \left[\log \left(\left| \sum_{k=1}^{n_t} |g_{ik}| \cos(\Theta_{i,k} + \tilde{\Phi}_{ik}) x_k \right|^2 \right) \right] \\ &\quad + \log \|\tilde{\mathbf{X}}\| \\ &\geq \log \|\tilde{\mathbf{X}}\| + c_H \quad (48) \\ &\geq \log |\tilde{X}_U| + c_H, \quad (49) \end{aligned}$$

where $\tilde{\Phi}_{ik} \triangleq \angle_{g_{ik} \tilde{X}_k}$; (48) is obtained by applying Lemma 6 with $\mathbf{V} \triangleq \left[|g_{ik}| \cos(\Theta_{T,k} + \tilde{\Phi}_{ik}) \right]_k$ with $\mathbb{E}[\|\mathbf{V}\|^2] \leq \sum_k |g_{ik}|^2 < \infty$ and

$$\begin{aligned} h(\mathbf{V}) &\geq h(\mathbf{V} | \tilde{\Phi}) \\ &= h(\cos(\Theta_T + \tilde{\Phi}) | \tilde{\Phi}) + \sum_k \log |g_{ik}| \\ &> -\infty, \end{aligned}$$

where the equality is the application of the change of variables from Lemma 2; the last inequality is from Lemma 5. From (47) and (49), we get

$$\begin{aligned} h(W_i | \tilde{\mathbf{X}}, W^{i-1}) &= \mathbb{E}_{\tilde{\mathbf{X}}} \left[\mathbb{E}[\log^+ |B_i| | \tilde{\mathbf{X}}] \right] + c_0 \\ &\geq \mathbb{E}_{\tilde{\mathbf{X}}} \left[\left(\mathbb{E}[\log |B_i| | \tilde{\mathbf{X}}] \right)^+ \right] + c_0 \\ &\geq \mathbb{E} \left[(\log |\tilde{X}_U| + c_H)^+ \right] + c_0 \\ &\geq \mathbb{E}[\log^+ |\tilde{X}_U|] - |c_H| + c_0, \end{aligned} \quad (50)$$

where the last inequality is from the application of (12) in Lemma 8 with $p = 2^{c_H}$. Plugging (45) and (50) into (44), the lower bound (25) is obtained.

C. Proof of Lemma 10

From Definition 1, by imposing $1 > \alpha_i > 0$, $i = 1, \dots, n_r$, and $\mu = \min\{P^{-1}, 1\}$, we have

$$\begin{aligned} -\log q(\mathbf{W}) &= -\log \frac{g\alpha}{n_r!} - \sum_{i=1}^{n_r} \alpha_i \log \mu + \mu |\hat{W}_{n_r}|^2 \\ &\quad + \sum_{i=2}^{n_r} (1 - \alpha_i) \log (|\hat{W}_i|^2 - |\hat{W}_{i-1}|^2) \\ &\quad + (1 - \alpha_1) \log |\hat{W}_1|^2. \end{aligned} \quad (51)$$

We bound each term as follows.

- The squared magnitude of each output

$$\begin{aligned} |W_i|^2 &\leq 2|\mathbf{G}_i^T \tilde{\mathbf{X}}|^2 + 2|Z_i|^2 \\ &\leq 2\|\mathbf{G}_i\|^2 \|\tilde{\mathbf{X}}\|^2 + 2|Z_i|^2 \\ &\leq 2\lambda_{\mathbf{H}} \|\tilde{\mathbf{X}}\|^2 + 2\|\mathbf{Z}\|^2, \end{aligned} \quad (52)$$

where \mathbf{G}_i^T is the i th row of the canonical matrix $\mathbf{G}_{(U)}$ defined in (15); (52) is due to Cauchy-Schwarz; and $\lambda_{\mathbf{H}}$ is defined as

$$\lambda_{\mathbf{H}} \triangleq \max_{u=1, \dots, n_t} \|\mathbf{G}_{(u)}\|^2.$$

- The difference of the squared magnitudes

$$\begin{aligned} &||W_i|^2 - |W_k|^2| \\ &\leq (|W_i| + |W_k|) |e^{-j\Theta_{i,U}} W_i - e^{-j\Theta_{k,U}} W_k| \\ &\leq 2^{\frac{3}{2}} \sqrt{\lambda_{\mathbf{H}} \|\tilde{\mathbf{X}}\|^2 + \|\mathbf{Z}\|^2} |e^{-j\Theta_{i,U}} W_i - e^{-j\Theta_{k,U}} W_k| \end{aligned}$$

with

$$\begin{aligned} &|e^{-j\Theta_{i,U}} W_i - e^{-j\Theta_{k,U}} W_k|^2 \\ &\leq \left| e^{-j\Theta_{i,U}} (\mathbf{G}_i^T \tilde{\mathbf{X}} + Z_i) - e^{-j\Theta_{k,U}} (\mathbf{G}_k^T \tilde{\mathbf{X}} + Z_k) \right|^2 \\ &\leq 2 \left| e^{-j\Theta_{i,U}} \mathbf{G}_i^T \tilde{\mathbf{X}} - e^{-j\Theta_{k,U}} \mathbf{G}_k^T \tilde{\mathbf{X}} \right|^2 + 2|Z_i|^2 + 2|Z_k|^2 \\ &\leq 2 \left(\sum_{l \neq U} |e^{-j\Theta_{i,U}} G_{il} - e^{-j\Theta_{k,U}} G_{kl}|^2 \right) \left(\sum_{l \neq U} |\tilde{X}_l|^2 \right) \\ &\quad + 2\|\mathbf{Z}\|^2 \\ &\leq 4(n_t - 1) \lambda_{\mathbf{H}} \tilde{X}_V^2 + 2\|\mathbf{Z}\|^2. \end{aligned}$$

Note that the above upper bounds does not depend on i and k . Then, with the above bounds, we take expectation of the terms in (51), and obtain

$$\begin{aligned} \mathbb{E}[|\hat{W}_{n_r}|^2] &\leq 2(\lambda_{\mathbf{H}} \mathbb{E}[\|\tilde{\mathbf{X}}\|^2] + \mathbb{E}[\|\mathbf{Z}\|^2]) \\ &\leq 2\lambda_{\mathbf{H}} a^2 P + c_0, \end{aligned} \quad (53)$$

$$\begin{aligned} \mathbb{E}[\log |\hat{W}_1|^2] &\leq \mathbb{E}[\log(\lambda_{\mathbf{H}} \|\tilde{\mathbf{X}}\|^2 + \mathbb{E}[\|\mathbf{Z}\|^2])] + 1 \\ &\leq 2\mathbb{E}[\log^+ |\tilde{X}_U|] + c_H, \end{aligned} \quad (54)$$

where the last inequality is from Lemma 8. Similarly, basic calculations lead to

$$\begin{aligned} &\mathbb{E} \left[\log \left| |\hat{W}_i|^2 - |\hat{W}_{i-1}|^2 \right| \right] \\ &\leq \frac{1}{2} \mathbb{E} \left[\log \left(\lambda_{\mathbf{H}} \|\tilde{\mathbf{X}}\|^2 + \mathbb{E}[\|\mathbf{Z}\|^2] \right) \right] \\ &\quad + \frac{1}{2} \mathbb{E} \left[\log \left(4(n_t - 1) \lambda_{\mathbf{H}} |\tilde{X}_V|^2 + 2\mathbb{E}[\|\mathbf{Z}\|^2] \right) \right] + c_0 \\ &\leq \log^+ |\tilde{X}_U| + \log^+ |\tilde{X}_V| + c_H. \end{aligned} \quad (55)$$

Taking expectation over \mathbf{X} in (51), and plugging (53), (54), and (55) into it, we readily obtain (22).

D. Proof of Lemma 13

To prove Lemma 13, we deal with the cases $n_r = 2n_t - 1$ and $n_r \neq 2n_t - 1$ separately. Let us define $\hat{\mathbf{Y}}$ and $\tilde{\mathbf{Y}}$ such that

$$\mathbf{Y} = \sqrt{\frac{P}{n_t}} e^{j\tilde{\Theta}_R} \circ \hat{\mathbf{Y}} + \mathbf{Z} = \sqrt{\frac{P}{n_t}} \tilde{\mathbf{Y}} + \mathbf{Z}.$$

For notational convenience, we define $n \triangleq n_r$ and $m \triangleq n_t$ in the following proof.

1) *Case $n = 2m - 1$.* First we show that (34) holds for $n = 2m - 1$. We write $h(\mathbf{Y}) \geq h(\mathbf{Y} | \mathbf{Z}) = h(\mathbf{Y} - \mathbf{Z} | \mathbf{Z}) = h\left(\sqrt{\frac{P}{m}} \tilde{\mathbf{Y}}\right) = n \log P + h(\tilde{\mathbf{Y}}) + c_0$. Now, it is enough to show that $h(\tilde{\mathbf{Y}}) > -\infty$. From Lemma 4,

$$\begin{aligned} h(\tilde{\mathbf{Y}}) &\geq h(|\tilde{\mathbf{Y}}|^2) + c_0 \\ &= h(|\hat{\mathbf{Y}}|^2) + c_0 \\ &\geq h(\mathbf{S} | \hat{Y}_n) + h(|\hat{Y}_n|^2) + c_0 \\ &= h(\mathbf{S} | \hat{Y}_n) + c_H, \end{aligned}$$

where $\mathbf{S} \in \mathbb{R}^{n-1}$ with $S_i \triangleq |\hat{Y}_i|^2$ for $i = 1, \dots, n - 1$; the second inequality is from the chain rule and that adding the condition on the phase of \hat{Y}_n reduces entropy; the last equality is due to $\hat{Y}_n \sim \mathcal{CN}(0, m^{-1} \|\mathbf{h}_n\|^2)$. Next we need to show that $h(\mathbf{S} | \hat{Y}_n) > -\infty$. Intuitively, given \hat{Y}_n , \mathbf{S} can

be expressed as $n - 1 = 2(m - 1)$ real functions of the $2(m - 1)$ real random variables $(\text{Re}\{\hat{Y}^{m-1}\}, \text{Im}\{\hat{Y}^{m-1}\})$. Since $h(\text{Re}\{\hat{Y}^{m-1}\}, \text{Im}\{\hat{Y}^{m-1}\}) = h(\hat{Y}^{m-1})$ is finite for almost every \mathbf{H} , as long as the mapping is not *degenerated*, $h(\mathbf{S} | \hat{Y}_n)$ should be finite too. This argument is proved formally in the following.

Since for any generic $\mathbf{H} \in \mathbb{C}^{n \times m}$, any m rows of the matrix are linear independent, the remaining $n - m$ rows can be written as linear combinations of these rows. Let us take the rows $\{1, 2, \dots, m - 1, n\}$ of \mathbf{H} . It readily follows that one can write

$$\hat{Y}_m^{n-1} = \mathbf{B}\hat{Y}^{m-1} + \mathbf{b}\hat{Y}_n$$

with $\mathbf{B} \in \mathbb{C}^{(m-1) \times (m-1)}$ and $\mathbf{b} \in \mathbb{C}^{(m-1) \times 1}$ depending only on \mathbf{H} . Next let us partition the space \mathbb{R}^{n-1} into a finite number of sets in each one of which the mapping $(\text{Re}\{\hat{y}^{m-1}\}, \text{Im}\{\hat{y}^{m-1}\}) \mapsto \mathbf{s}$ is bijective. Note that this is possible since the mapping is quadratic in a finite-dimensional space. Let Ω be the index of the partitions which only depends on \hat{Y}^{m-1} . Then

$$\begin{aligned} h(\mathbf{S} | \hat{Y}_n) &\geq h(\mathbf{S} | \hat{Y}_n, \Omega) \\ &= \mathbb{E}[\log |\det(\mathbf{J})|] + h(\hat{Y}^{m-1} | \hat{Y}_n, \Omega) \\ &= \mathbb{E}[\log |\det(\mathbf{J})|] + c_H, \end{aligned}$$

where the first equality is from Lemma 2; the second equality is due to $h(\hat{Y}^{m-1} | \Omega, \hat{Y}_n) = h(\hat{Y}^{m-1} | \hat{Y}_n) - I(\Omega; \hat{Y}^{m-1} | \hat{Y}_n)$ with $h(\hat{Y}^{m-1} | \hat{Y}_n) > -\infty$ for any generic \mathbf{H} and $I(\Omega; \hat{Y}^{m-1} | \hat{Y}_n) \leq H(\Omega) < \infty$; \mathbf{J} is the Jacobian matrix with

$$\begin{aligned} \det(\mathbf{J}) &= \det \begin{bmatrix} \frac{\partial \mathbf{S}}{\partial \text{Re}\{\hat{Y}^{m-1}\}} & \frac{\partial \mathbf{S}}{\partial \text{Im}\{\hat{Y}^{m-1}\}} \\ \frac{\partial \mathbf{S}}{\partial \hat{Y}_n} & \frac{\partial \mathbf{S}}{\partial \hat{Y}_n} \end{bmatrix} \\ &= 2^{m-1} \det \begin{bmatrix} \frac{\partial \mathbf{S}}{\partial \hat{Y}^{m-1}} & \frac{\partial \mathbf{S}}{\partial (\hat{Y}^{m-1})^*} \\ \frac{\partial \mathbf{S}}{\partial \hat{Y}_n} & \frac{\partial \mathbf{S}}{\partial \hat{Y}_n} \end{bmatrix} \quad (56) \\ &= 2^{m-1} \det \begin{bmatrix} \text{diag}\{(\hat{Y}^{m-1})^*\} & \text{diag}\{\hat{Y}^{m-1}\} \\ \text{diag}\{(\hat{Y}_m^{n-1})^*\}\mathbf{B} & \text{diag}\{\hat{Y}_m^{n-1}\}\mathbf{B}^* \end{bmatrix} \\ &= 4^{m-1} j^{m-1} \text{Im} \left\{ \text{diag}\{\hat{Y}_m^{n-1}\}\mathbf{B}^* \text{diag}\{(\hat{Y}^{m-1})^*\} \right\}, \quad (57) \end{aligned}$$

where (56) is due to the fact that the complex gradient of a real-valued function is a unitary transformation of the real gradient (see, e.g., [21, App.A6]); to obtain the last equality, we apply the identity $\det \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} = \det(\mathbf{C})\det(\mathbf{F} - \mathbf{E}\mathbf{C}^{-1}\mathbf{D})$. Since $\hat{Y}_1, \dots, \hat{Y}_{m-1}, \hat{Y}_n$ are jointly circularly symmetric Gaussian with finite and non-degenerate covariance for any generic \mathbf{H} , there exists a \hat{Y}_n' circularly symmetric with non-zero bounded variance and independent of \hat{Y}^{m-1} , such that

$$\begin{aligned} \hat{Y}_n &= \hat{Y}_n' + \mathbf{f}^\top \hat{Y}^{m-1}, \quad \text{and thus} \\ \hat{Y}_m^{n-1} &= (\mathbf{B} + \mathbf{b}\mathbf{f}^\top)\hat{Y}^{m-1} + \mathbf{b}\hat{Y}_n', \end{aligned}$$

where $\mathbf{f} \in \mathbb{C}^{(m-1) \times 1}$ depends only on \mathbf{H} . Then we can continue from (57) and write $|\det(\mathbf{J})|$ as

$$\begin{aligned} |\det(\mathbf{J})| &= 4^{m-1} \left| \det \left(\text{Im}\{\hat{Y}_n' \text{diag}\{\mathbf{b}\}\mathbf{A}\} + \mathbf{M} \right) \right| \\ &= 4^{m-1} \left| \det \left(\hat{Y}_{n,R}' \mathbf{N}_I + \hat{Y}_{n,I}' \mathbf{N}_R + \mathbf{M} \right) \right|, \end{aligned}$$

where $\mathbf{A} \triangleq \mathbf{B}^* \text{diag}\{(\hat{Y}^{m-1})^*\}$ and $\mathbf{M} \triangleq \text{Im}\left\{ \text{diag}\{(\mathbf{B} + \mathbf{b}\mathbf{f}^\top)\hat{Y}^{m-1}\}\mathbf{B}^* \text{diag}\{(\hat{Y}^{m-1})^*\} \right\}$; $\hat{Y}_{n,R}'$ and $\hat{Y}_{n,I}'$ are the real and imaginary parts of \hat{Y}_n' , respectively; \mathbf{N}_R and \mathbf{N}_I are the real and imaginary parts of $\text{diag}\{\mathbf{b}\}\mathbf{A}$, respectively. Conditioned on \hat{Y}^{m-1} and $\hat{Y}_{n,I}'$, the determinant $|\det(\mathbf{J})|$ can be written as the absolute value of the characteristic polynomial of $(\hat{Y}_{m,I}' \mathbf{N}_R + \mathbf{M})\mathbf{N}_I^{-1}$, namely,

$$|\det(\mathbf{J})| = 4^{m-1} |\det(\mathbf{N}_I)| \prod_{t=1}^{m-1} |\hat{Y}_{n,R}' - \Lambda_t|,$$

where $\Lambda_1, \dots, \Lambda_{m-1}$ are the eigenvalues of $(\hat{Y}_{n,I}' \mathbf{N}_R + \mathbf{M})\mathbf{N}_I^{-1}$ in \mathbb{C} and are functions of \mathbf{H} , \hat{Y}^{m-1} , and $\hat{Y}_{n,I}'$. Then

$$\begin{aligned} \mathbb{E} \left[\log |\det(\mathbf{J})| | \hat{Y}^{m-1}, \hat{Y}_{n,I}' \right] &= \mathbb{E} \left[\log |\det(\mathbf{N}_I)| | \hat{Y}^{m-1}, \hat{Y}_{n,I}' \right] \\ &\quad + \sum_{t=1}^{m-1} \mathbb{E} \left[\log |\hat{Y}_{n,R}' - \Lambda_t| | \hat{Y}^{m-1}, \hat{Y}_{n,I}' \right] + c_0 \\ &\geq \mathbb{E} \left[\log |\det(\mathbf{N}_I)| | \hat{Y}^{m-1}, \hat{Y}_{n,I}' \right] + \sum_{t=1}^{m-1} \mathbb{E} \left[\log |\hat{Y}_{n,R}'| \right] + c_0 \\ &= \mathbb{E} \left[\log |\det(\mathbf{N}_I)| | \hat{Y}^{m-1}, \hat{Y}_{n,I}' \right] + c_H, \end{aligned}$$

where the inequality is from $\log |\hat{Y}_{n,R}' - \Lambda_t| \geq \log |\hat{Y}_{n,R}' - \text{Re}\{\Lambda_t\}|$, the application of Lemma 1 with $k = 1$, and the independence between $\hat{Y}_{n,R}'$ and $(\hat{Y}^{m-1}, \hat{Y}_{n,I}')$. Thus, taking expectation over $(\hat{Y}^{m-1}, \hat{Y}_{n,I}')$, we have

$$\mathbb{E}[\log |\det(\mathbf{J})|] \geq \mathbb{E}[\log |\det(\mathbf{N}_I)|] + c_H.$$

It remains to show that $\mathbb{E}[\log |\det(\mathbf{N}_I)|] > -\infty$. Let us recall that $\mathbf{N}_I \triangleq \text{Im}\{\text{diag}\{\mathbf{b}\}\mathbf{B}^* \text{diag}\{(\hat{Y}^{m-1})^*\}\} = \mathbf{T}_I \text{diag}\{\hat{Y}_R^{m-1}\} - \mathbf{T}_R \text{diag}\{\hat{Y}_I^{m-1}\}$ where \mathbf{T}_R and \mathbf{T}_I are the real and imaginary parts of $\text{diag}\{\mathbf{b}\}\mathbf{B}^*$, respectively, and thus depends only on \mathbf{H} . Then

$$\begin{aligned} \mathbb{E}[\log |\det(\mathbf{N}_I)|] &= \mathbb{E}[\log |\det(\mathbf{T}_I)|] \\ &= \mathbb{E} \left[\log \left| \det \left(\text{diag}\{\hat{Y}_R^{m-1}\} - \mathbf{T}_I^{-1} \mathbf{T}_R \text{diag}\{\hat{Y}_I^{m-1}\} \right) \right| \right] \quad (58) \end{aligned}$$

We shall show that (58) is lower-bounded as follows. Note that $\hat{Y}_{m-1,R}$ can be written as $\hat{Y}_{m-1,R} = \hat{Y}'_{m-1,R} + L_{m-1}(\hat{Y}_R^{m-2}, \hat{Y}_I^{m-1})$ with $\hat{Y}'_{m-1,R}$ centered normal with non-zero bounded variance and being independent of $(\hat{Y}_R^{m-2}, \hat{Y}_I^{m-1})$, and L_{m-1} some linear operator that depends only on \mathbf{H} . Let $\mathbf{P}_{m-1} \triangleq \mathbf{T}_I^{-1} \mathbf{T}_R$. It is easy to verify that $\det(\text{diag}\{\hat{Y}_R^{m-1}\} - \mathbf{P}_{m-1} \text{diag}\{\hat{Y}_I^{m-1}\})$ can be written as $\det(\text{diag}\{\hat{Y}_R^{m-2}\} - \mathbf{P}_{m-2} \text{diag}\{\hat{Y}_I^{m-2}\})\hat{Y}'_{m-1,R} + L'_{m-1}(\hat{Y}_R^{m-2}, \hat{Y}_I^{m-1})$ where \mathbf{P}_{m-2} is the upper-left $(m-2) \times (m-2)$ sub-matrix of \mathbf{P}_{m-1} and $L'_{m-1}(\hat{Y}_R^{m-2}, \hat{Y}_I^{m-1})$ is some value that is independent of $\hat{Y}'_{m-1,R}$. Thus, we can again apply

Lemma 1 and obtain recursively

$$\begin{aligned}
& \mathbb{E} \left[\log \left| \det \left(\text{diag} \{ \hat{Y}_R^{m-1} \} - P_{m-1} \text{diag} \{ \hat{Y}_I^{m-1} \} \right) \right| \right] \\
& \geq \mathbb{E} \left[\log \left| \det \left(\text{diag} \{ \hat{Y}_R^{m-2} \} - P_{m-2} \text{diag} \{ \hat{Y}_I^{m-2} \} \right) \hat{Y}'_{m-1, R} \right| \right] \\
& \vdots \\
& \geq \mathbb{E} \left[\log \left| \hat{Y}_{1, R} - P_{1,1} \hat{Y}_{1, I} \right| \right] + \sum_{k=1}^{m-1} \mathbb{E} \left[\log \left| \hat{Y}'_{k, R} \right| \right] \\
& \geq c_H,
\end{aligned}$$

where the last inequality is from $\mathbb{E} \left[\log \left| \hat{Y}_{1, R} - P_{1,1} \hat{Y}_{1, I} \right| \right] \geq \mathbb{E} \left[\log \left| \hat{Y}_{1, R} \right| \right] \geq c_H$ due to the independence between $\hat{Y}_{1, R}$ and $\hat{Y}_{1, I}$ and the application of Lemma 1.

Finally, recalling that $\mathbf{T}_I \triangleq \text{Im} \{ \text{diag} \{ \mathbf{b} \} \mathbf{B}^* \}$, we have $\log |\det(\mathbf{T}_I)| > -\infty$ for any generic \mathbf{H} , it follows from (58) that $\mathbb{E} \left[\log |\det(\mathbf{N}_I)| \right]$ is lower-bounded. By now, we have shown that $h(\mathbf{Y}) \geq n \log P + c_H$. Since $h(\mathbf{Y}) \geq h(\mathbf{Y} | \tilde{\mathbf{Y}}) = h(\mathbf{Z}) \geq n \log(\pi e) \geq 0$, we have $h(\mathbf{Y}) \geq (n \log P + c_H)^+ \geq n \log^+ P - |c_H|$ from (12) in Lemma 8 with $p = \frac{c_H}{n}$. This completes the proof for the case $n = 2m - 1$.

2) *Case $n \neq 2m - 1$* : Note that if (34) holds for $n = 2m - 1$, then it also holds for $n < 2m - 1$ and $n > 2m - 1$. To see this, in the case with $n < 2m - 1$, we can add $2m - 1 - n$ receive antennas to have $(\mathbf{Y}, \mathbf{Y}')$ with \mathbf{Y}' being the extra outputs. Since (34) holds for $h(\mathbf{Y}, \mathbf{Y}')$ by assumption, then we have

$$\begin{aligned}
h(\mathbf{Y}) & \geq h(\mathbf{Y}, \mathbf{Y}') - h(\mathbf{Y}') \\
& \geq (2m - 1) \log^+ P - (2m - 1 - n) \log^+ P + c_H \\
& = n \log^+ P + c_H,
\end{aligned}$$

where the second inequality is from (34) and the fact that $h(\mathbf{Y}') \leq (2m - 1 - n) \log^+ P + c_H$. When $n > 2m - 1$, we partition $\mathbf{Y} = (\mathbf{Y}', \mathbf{Y}'')$ with $\mathbf{Y}' \in \mathbb{C}^{(2m-1) \times 1}$ and obtain

$$\begin{aligned}
h(\mathbf{Y}) & = h(\mathbf{Y}') + \sum_{k=2m}^n h(Y_k | Y^{k-1}) \\
& \geq (2m - 1) \log^+ P \\
& \quad + \sum_{k=2m}^n h(Y_k | Y^{k-1}, \tilde{\Theta}_R^{k-1}, \tilde{\Theta}_T, \mathbf{X}) + c_H \\
& \geq (2m - 1) \log^+ P \\
& \quad + \sum_{k=2m}^n \mathbb{E} \left[\log^+ \left| \sqrt{\frac{P}{m}} \mathbf{h}_k^T (e^{j\tilde{\Theta}_T} \circ \mathbf{X}_0) \right| \right] + c'_H \quad (59) \\
& = \left(2m - 1 + \frac{n - 2m + 1}{2} \right) \log^+ P + c''_H \\
& = \left(m + \frac{n - 1}{2} \right) \log^+ P + c''_H,
\end{aligned}$$

where the second inequality is from Lemma 7; the equality (59) is from the fact that $\mathbf{h}_k^T (e^{j\tilde{\Theta}_T} \circ \mathbf{X}_0) \sim \mathcal{CN}(0, \|\mathbf{h}_k\|^2)$.

ACKNOWLEDGEMENT

S. Yang would like to thank G. Durisi for helpful discussions and comments during the early stage of this work.

REFERENCES

- [1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless personal communications*, vol. 6, no. 3, pp. 311–335, 1998.
- [2] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European transactions on telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.
- [3] A. Lapidoth and S. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, Oct 2003.
- [4] L. Zheng and D. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Theory*, vol. 48, no. 2, pp. 359–383, Feb 2002.
- [5] A. Lapidoth, "On phase noise channels at high SNR," in *Proc. IEEE Information theory workshop*, Oct 2002, pp. 1–4.
- [6] M. Katz and S. Shamai, "On the capacity-achieving distribution of the discrete-time noncoherent and partially coherent AWGN channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2257–2270, Oct 2004.
- [7] L. Barletta, M. Magarini, and A. Spalvieri, "The information rate transferred through the discrete-time Wiener's phase noise channel," *Journal of Lightwave Technology*, vol. 30, no. 10, pp. 1480–1486, 2012.
- [8] H. Ghozlan and G. Kramer, "Models and information rates for Wiener phase noise channels," *arXiv preprint arXiv:1503.03130*, 2015.
- [9] G. Durisi, A. Tarable, C. Camarda, R. Devassy, and G. Montorsi, "Capacity bounds for MIMO microwave backhaul links affected by phase noise," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 920–929, March 2014.
- [10] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavčić, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3498–3508, 2006.
- [11] M. R. Khanzadi, G. Durisi, and T. Eriksson, "Capacity of SIMO and MISO phase-noise channels with common/separate oscillators," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3218–3231, 2015.
- [12] G. Durisi, A. Tarable, and T. Koch, "On the multiplexing gain of MIMO microwave backhaul links affected by phase noise," in *Proc. IEEE International Conference on Commun. (ICC)*, 2013, pp. 3209–3214.
- [13] A. A. Nasir, H. Mehrpouyan, and R. Schober, "Phase noise in MIMO systems: Bayesian Cramér-Rao bounds and soft-input estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2675–2692, 2013.
- [14] T. Datta and S. Yang, "Improving MIMO detection performance in presence of phase noise using norm difference criterion," in *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 286–292.
- [15] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Math. Hung.*, vol. 2, pp. 291–292, 1967.
- [16] J. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Mathematicae*, vol. 77, no. 2, pp. 101–115, 1974.
- [17] H. Ghozlan and G. Kramer, "On Wiener phase noise channels at high signal-to-noise ratio," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, 2013, pp. 2279–2283.
- [18] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [19] A. M. Mathal and P. G. Moschopoulos, "A form of multivariate Gamma distribution," *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 97–106, March 1992.
- [20] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982.
- [21] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall Upper Saddle River, NJ, 2000.

Sheng Yang (M'07) received the B.E. degree in electrical engineering from Jiaotong University, Shanghai, China, in 2001, and both the engineer degree and the M.Sc. degree in electrical engineering from Telecom ParisTech, Paris, France, in 2004, respectively. In 2007, he obtained his Ph.D. from Université de Pierre et Marie Curie (Paris VI). From October 2007 to November 2008, he was with Motorola Research Center in Gif-sur-Yvette, France, as a senior staff research engineer. Since December 2008, he has joined CentraleSupélec where he is currently an associate professor. From April 2015, he also holds an honorary faculty position in the department of electrical and electronic engineering of the University of Hong Kong (HKU). He received the 2015 IEEE ComSoc Young Researcher Award for the Europe, Middle East, and Africa Region (EMEA). He is an editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Shlomo Shamai (Shitz) (F'94) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Technion-Israel Institute of Technology, in 1975, 1981 and 1986 respectively.

During 1975-1985 he was with the Communications Research Labs, in the capacity of a Senior Research Engineer. Since 1986 he is with the Department of Electrical Engineering, Technion-Israel Institute of Technology, where he is now a Technion Distinguished Professor, and holds the William Fondiller Chair of Telecommunications. His research interests encompasses a wide spectrum of topics in information theory and statistical communications.

Dr. Shamai (Shitz) is an IEEE Fellow, a member of the Israeli Academy of Sciences and Humanities and a foreign member of the US National Academy of Engineering. He is the recipient of the 2011 Claude E. Shannon Award and the 2014 Rothschild Prize in Mathematics/Computer Sciences and Engineering.

He has been awarded the 1999 van der Pol Gold Medal of the Union Radio Scientifique Internationale (URSI), and is a co-recipient of the 2000 IEEE Donald G. Fink Prize Paper Award, the 2003, and the 2004 joint IT/COM societies paper award, the 2007 IEEE Information Theory Society Paper Award, the 2009 and 2015 European Commission FP7, Network of Excellence in Wireless COMMunications (NEWCOM++, NEWCOM#) Best Paper Awards, the 2010 Thomson Reuters Award for International Excellence in Scientific Research, the 2014 EURASIP Best Paper Award (for the EURASIP Journal on Wireless Communications and Networking), and the 2015 IEEE Communications Society Best Tutorial Paper Award. He is also the recipient of 1985 Alon Grant for distinguished young scientists and the 2000 Technion Henry Taub Prize for Excellence in Research. He has served as Associate Editor for the Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY, and has also served twice on the Board of Governors of the Information Theory Society. He has served on the Executive Editorial Board of the IEEE TRANSACTIONS ON INFORMATION THEORY.