Coherent multiple-antenna block-fading channels at finite blocklength

Austin Collins and Yury Polyanskiy

Abstract

In this paper we consider a channel model that is often used to describe the mobile wireless scenario: multiple-antenna additive white Gaussian noise channels subject to random (fading) gain with full channel state information at the receiver. Dynamics of the fading process are approximated by a piecewise-constant process (frequency non-selective isotropic block fading). This work addresses the finite blocklength fundamental limits of this channel model. Specifically, we give a formula for the channel dispersion – a quantity governing the delay required to achieve capacity. The multiplicative nature of the fading disturbance leads to a number of interesting technical difficulties that required us to enhance traditional methods for finding the channel dispersion. Alas, one difficulty remains: the converse (impossibility) part of our result holds under an extra constraint on the growth of the peak-power with blocklength.

Our results demonstrate, for example, that while capacities of $n_t \times n_r$ and $n_r \times n_t$ antenna configurations coincide (under fixed received power), the coding delay can be sensitive to this switch. For example, at the received SNR of 20 dB the 16×100 system achieves capacity with codes of length (delay) which is only 60% of the length required for the 100×16 system. Another interesting implication is that for the MISO channel, the dispersion-optimal coding schemes require employing orthogonal designs such as Alamouti's scheme – a surprising observation considering the fact that Alamouti's scheme was designed for reducing demodulation errors, not improving coding rate. Finding these dispersion-optimal coding schemes naturally gives a criteria for producing orthogonal design-like inputs in dimensions where orthogonal designs do not exist.

I. INTRODUCTION

Given a noisy communication channel, the maximal cardinality of a codebook of blocklength n which can be decoded with block error probability no greater than ϵ is denoted as $M^*(n, \epsilon)$. The evaluation of this function – the fundamental performance limit of block coding – is alas computationally impossible for most channels of interest. As a resolution of this difficulty [1] proposed a closed-form normal approximation, based on the asymptotic expansion:

$$\log M^*(n, \epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n), \tag{1}$$

where the capacity C and dispersion V are two intrinsic characteristics of the channel and $Q^{-1}(\epsilon)$ is the inverse of the Q-function[1]. One immediate consequence of the normal approximation is an estimate for the minimal blocklength (delay) required to achieve a given fraction η of the channel capacity:

$$n \gtrsim \left(\frac{Q^{-1}(\epsilon)}{1 - \eta}\right)^2 \frac{V}{C^2}. \tag{2}$$

Asymptotic expansions such as (1) are rooted in the central-limit theorem and have been known classically for discrete memoryless channels [2], [3] and later extended in a wide variety of directions; see the surveys in [4], [5].

[1]As usual, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$.

The fading channel is the centerpiece of the theory and practice of wireless communication, and hence there are many slightly different variations of the model: differing assumptions on the dynamics and distribution of the fading process, antenna configurations, and channel state knowledge. The capacity of the fading channel was found independently by Telatar [6] and Foschini and Gans [7] for the case of Rayleigh fading and channel state information available at the receiver only (CSIR) and at both the transmitter and receiver (CSIRT). Motivated by the linear gains promised by capacity results, space time codes were introduced to exploit multiple antennas, most notable amongst them is Alamouti's ingenious orthogonal scheme [8] along with a generalization of Tarokh, Jafarkhani and Calderbank [9]. Motivated by a recent surge of orthogonal frequency division (OFDM) technology, this paper focuses on an isotropic channel gain distribution, which is piecewise independent ("block-fading") and assume full channel state information available at the receiver (CSIR). This work describes finite blocklength effects incurred by the fading on the fundamental communication limits.

Some of the prior work on similar questions is as follows. Single antenna channel dispersion was computed in [10] for a more general stationary channel gain process with memory. In [11] finite-blocklength effects are explored for the non-coherent block fading setup. Quasi-static fading channels in the general MIMO setting have been thoroughly investigated in [12], showing that the expansion (1) changes dramatically (in particular the channel dispersion term becomes zero); see also [13] for evaluation of the bounds. Coherent quasi-static channel has been studied in the limit of infinitely many antennas in [14] appealing to concentration properties of random matrices. Dispersion for lattices (infinite constellations) in fading channels has been investigated in a sequence of works, see [15] and references. Note also that there are some very fine differences between stationary and block-fading channel models, cf. [16, Section 4]. The minimum energy to send k bits over a MIMO channel for both the coherent and non-coherent case was studied in [17], showing the latter requires orders of magnitude larger latencies. [18] investigates the problem of power control with an average power constraint on the codebook in the quasi-static fading channel with perfect CSIRT. A novel achievability bound was found and evaluated for the fading channel with CSIR in [19]. Parts of this work have previously appeared in [20], [21].

The paper is organized as follows. In Section II we describe the channel model and state all our main results formally. Section III characterizes capacity achieving input/output distributions (caid/caod, resp.) and evaluates moments of the information density. Then in Sections IV and V we prove the achievability and converse parts of our (non rank-1) results, respectively. Section VI focuses on the special case of when the matrix of channel gains has rank 1. Finally, Section VII contains a discussion of numerical results and the behavior of channel dispersion as a function of the number of antennas.

The numerical software used to compute the achievability bounds, dispersion and normal approximation in this work can be found online under the Spectre project [22].

II. Main Results

A. Channel Model

The channel model considered in this paper is the frequency-nonselective coherent real block fading (BF) discrete-time channel with multiple transmit and receive antennas (MIMO) (See [23, Section II] for background on this model). We will simply refer to it as the *MIMO-BF channel*, which we formally define here. Given parameters n_t, n_r, P, T as follows: let $n_t \geq 1$ be the number of transmit antennas, $n_r \geq 1$ be the number of receive antennas, and $T \geq 1$ be the coherence time of the channel. The input-output relation at block j (spanning time instants $(j-1)T + 1$ to jT) with $j = 1, \ldots, n$ is given by

$$Y_j = H_j X_j + Z_j, \tag{3}$$

where $\{H_j, j = 1, \ldots\}$ is a $n_r \times n_t$ matrix-valued random fading process, X_j is a $n_t \times T$ matrix channel input, Z_j is a $n_r \times T$ Gaussian random real-valued matrix with independent entries of zero mean and unit variance, and Y_j is the $n_r \times T$ matrix-valued channel output. The process H_j is assumed to be i.i.d. with

isotropic distribution P_H, i.e. for any orthogonal matrices $U \in \mathbb{R}^{n_r \times n_r}$ and $V \in \mathbb{R}^{n_t \times n_t}$, both UH and HV are equal in distribution to H. We also assume

$$\mathbb{P}[H \neq 0] > 0 \tag{4}$$

to avoid trivialities. Note that due to merging channel inputs at time instants $1, \ldots, T$ into one matrix-input, the block-fading channel becomes memoryless. We assume coherent demodulation so that the channel state information (CSI) H_j is fully known to the receiver (CSIR).

An $(nT, M, \epsilon, P)_{CSIR}$ code of blocklength nT, probability of error ϵ and power-constraint P is a pair of maps: the encoder $f : [M] \to (\mathbb{R}^{n_t \times T})^n$ and the decoder $g : (\mathbb{R}^{n_r \times T})^n \times (\mathbb{R}^{n_r \times n_t})^n \to [M]$ satisfying the probability of error constraint

$$\mathbb{P}[W \neq \hat{W}] \leq \epsilon . \tag{5}$$

on the probability space

$$W \to X^n \to (Y^n, H^n) \to \hat{W} ,$$

where the message W is uniformly distributed on $[M]$, $X^n = f(W)$, $X^n \to (Y^n, H^n)$ is as described in (3), and $\hat{W} = g(Y^n, H^n)$. In addition the input sequences are required to satisfy the power constraint:

$$\sum_{j=1}^{n} \|X_j\|_F^2 \leq nTP \qquad \mathbb{P}\text{-a.s.} ,$$

where $\|M\|_F^2 \overset{\triangle}{=} \sum_{i,j} M_{i,j}^2$ is the Frobenius norm of the matrix M.

Under the isotropy assumption on P_H, the capacity C appearing in (1) of this channel is given by [6]

$$C(P) = \frac{1}{2}\mathbb{E}\left[\log \det\left(I_{n_r} + \frac{P}{n_t}HH^T\right)\right] \tag{6}$$

$$= \sum_{i=1}^{n_{\min}} \mathbb{E}\left[C_{AWGN}\left(\frac{P}{n_t}\Lambda_i^2\right)\right] , \tag{7}$$

where $C_{AWGN}(P) = \frac{1}{2}\log(1 + P)$ is the capacity of the additive white Gaussian noise (AWGN) channel with SNR P, $n_{\min} = \min(n_r, n_t)$ is the minimum of the transmit and receive antennas, and $\{\Lambda_i^2, i = 1, \ldots, n_{\min}\}$ are eigenvalues of HH^T. Note that it is common to think that as $P \to \infty$ the expression (7) scales as $n_{\min} \log P$, but this is only true if $\mathbb{P}[\text{rank } H = n_{\min}] = 1$.

The goal of this line of work is to characterize the dispersion of the present channel. Since the channel is memoryless it is natural to expect, given the results in [1], [10], that dispersion (for $\epsilon < 1/2$) is given by

$$V(P) \overset{\triangle}{=} \inf_{P_X : I(X;Y|H)=C} \frac{1}{T}\mathbb{E}\text{Var}(i(X;Y,H)|X) \tag{8}$$

where we denoted (single T-block) information density by

$$i(x; y, h) \overset{\triangle}{=} \log \frac{dP_{Y,H|X=x}}{dP_{Y,H}^*}(y, h) \tag{9}$$

and $P_{Y,H}^*$ is the capacity achieving output distribution (caod). Justification of (8) as the actual (operational) dispersion, appearing in the expansion of $\log M^*(n, \epsilon)$ is by no means trivial and is the subject of this work.

B. Statement of Main Theorems

Here we formally state the main results, then go into more detail in the following sections. Our first result is an achievability and partial converse bound for the MIMO-BF fading channel for fixed parameters n_t, n_r, T, P.

Theorem 1. *For the MIMO-BF channel, there exists an $(nT, M, \epsilon, P)_{CSIR}$ maximal probability of error code with $0 < \epsilon < 1/2$ satisfying*

$$\log M \geq nTC(P) - \sqrt{nTV(P)}Q^{-1}(\epsilon) + o(\sqrt{n}) . \tag{10}$$

Furthermore, for any $\delta_n \to 0$ there exists $\delta_n' \to 0$ so that every $(nT, M, \epsilon, P)_{CSIR}$ code with extra constraint that $\max_j \|x^j\|_F \leq \delta_n n^{1/4}$, must satisfy

$$\log M \leq nTC(P) - \sqrt{nTV(P)}Q^{-1}(\epsilon) + \delta_n'\sqrt{n} \tag{11}$$

where the capacity $C(P)$ is given by (6) and dispersion $V(P)$ by (8).[2]

Proof. This follows from Theorem 16 and Theorem 19 below. □

Remark 1. Note that the converse has an extra constraint $\max_j \|x^j\|_F \leq \delta_n n^{1/4}$. Mathematically, this constraint is needed so that the n-fold information information density $i(x^n; Y^n, H^n)$ behaves Gaussian-like, via the Berry-Esseen theorem. For example, if x^n had $x_{11} = \sqrt{nTP}$ and zeroes in all other coordinates, then one term in the information density would $O(n)$ while the rest would be $O(1)$, and hence no asymptotic behavior would emerge. All known bounds to obtain the channel dispersion rely on approximating the information density by a Gaussian, and hence a fundamentally different method of analysis is needed to handle the situation where $\max_j \|x^j\|_F \geq \delta_n n^{1/4}$

Fortunately, this constraint is satisfied by most practical systems. To violate this constraint, a significant portion of the power budget must be poured into a single coherent block, which 1) creates a very large PAPR (often illegal by regulating bodies), and 2) does a poor job of exploiting the diversity gain from coding over multiple independent coherent blocks. Therefore, our results should give the behavior for any practical system.

Remark 2. The remainder term $o(\sqrt{n})$ in (11) depends on the hyper-parameters (n_t, n_r, P_H) in a complicated way, which we did not attempt to study here.

The behavior of dispersion found in Theorem 1 turns out to depend crucially on whether $\text{rank}(H) \leq 1$ a.s. or not. When $\text{rank}(H) > 1$, all capacity achieving input distributions (caids) yield the same conditional variance (8), yet when $\text{rank}(H) \leq 1$, the conditional variance varies over the set of caids. The following theorem discusses the case where $\mathbb{P}[\text{rank } H > 1] > 0$. In this case, the dispersion (8) can be calculated for the simplest Telatar caid (i.i.d. Gaussian matrix X). The following theorem gives full details.

Theorem 2. *Assume that $\mathbb{P}[\text{rank } H > 1] > 0$, then $V(P) = V_{iid}(P)$, where*

$$V_{iid}(P) = T\text{Var}\left(\sum_{i=1}^{n_{\min}} C_{AWGN}\left(\frac{P}{n_t}\Lambda_i^2\right)\right)$$

$$+ \sum_{i=1}^{n_{\min}} \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\Lambda_i^2\right)\right]$$

$$+ \left(\frac{P}{n_t}\right)^2\left(\eta_1 - \frac{\eta_2}{n_t}\right) \tag{12}$$

---

[2]For the explicit expression for $i(x; y, h)$ see (49) below.

where $\{\Lambda_i^2, i = 1, \ldots, n_{\min}\}$ *are eigenvalues of* HH^T, $V_{AWGN}(P) = \frac{\log^2 e}{2}\left(1 - \frac{1}{(1+P)^2}\right)$, *and*

$$c(\sigma) \triangleq \frac{\sigma}{1 + \frac{P}{n_t}\sigma} \tag{13}$$

$$\eta_1 \triangleq \frac{\log^2 e}{2} \sum_{i=1}^{n_{\min}} \mathbb{E}\left[c^2(\Lambda_i^2)\right] \tag{14}$$

$$\eta_2 \triangleq \frac{\log^2 e}{2} \left(\sum_{i=1}^{n_{\min}} \mathbb{E}\left[c(\Lambda_i^2)\right]\right)^2 \tag{15}$$

Proof. This is proved in Proposition 11 below. □

Remark 3. Each of the three terms in (12) is non-negative, see Remark 7 below for more details.

In the case where the fading process has rank 1 (e.g. for MISO systems), there are a multitude of caids, and the minimization problem in (8) is non-trivial. Quite surprisingly, for some values of n_t, T, we show that the (essentially unique) minimizer is a full-rate orthogonal design. The latter were introduced into the field of communications by Alamouti [8] and Tarokh et al [9]. This shows a somewhat unexpected connection between schemes optimal from modulation-theoretic and information-theoretic points of view. The precise results are as follows.

Theorem 3. *When* $\mathbb{P}[rank(H) \leq 1] = 1$, *we have*

$$V(P) = T\text{Var}\left(C_{AWGN}\left(\frac{P}{n_t}\Lambda^2\right)\right) + \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\Lambda^2\right)\right] \tag{16}$$

$$+ \left(\frac{P}{n_t}\right)^2\left(\eta_1 - \frac{\eta_2}{n_t^2 T}v^*(n_t, T)\right) \tag{17}$$

where Λ^2 *is the non-zero eigenvalues of* HH^T, *and*

$$v^*(n_t, T) = \frac{n_t^2}{2P^2} \max_{P_X : I(X;Y,H)=C} \text{Var}(\|X\|_F^2) \tag{18}$$

Proof. This is the content of Proposition 12 below. □

The quantity $v^*(n_t, T)$ is defined separately in Theorem 3 because it isolates how the dispersion depends on the input distribution. Unfortunately, $v^*(n_t, T)$ is generally unknown, since the maximization in (18) is over a manifold of matrix-valued random variables. However, for many dimensions, the maximum can be found by invoking the Hurwitz-Radon theorem [24]. We state this below to introduce the notation, and expand on it in Section VI.

Theorem 4 (Hurwitz-Radon). *There exists a family of* $n \times n$ *real matrices* $V_1, \ldots, V_k$ *satisfying*

$$V_i^T V_i = I_n \qquad i = 1, \ldots, k \tag{19}$$
$$V_i^T V_j + V_j^T V_i = 0 \quad i \neq j \tag{20}$$

if and only if $k \leq \rho(n)$, *where*

$$\rho(2^a b) = 8\left\lfloor\frac{a}{4}\right\rfloor + 2^{a \bmod 4}, \qquad a, b \in \mathbb{Z}, b\text{--}odd\,. \tag{21}$$

In particular, $\rho(n) \leq n$ *and* $\rho(n) = n$ *only for* $n = 1, 2, 4, 8$.

For a concrete example, note that Alamouti's scheme is created from a Hurwitz-Radon family for $n = k = 2$. Indeed, take the matrices

$$V_1 = I_2, \quad V_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

then Alamouti's orthogonal design can be formed by taking $aV_1 + bV_2$. It turns out that "maximal" Hurwitz-Radon families give capacity achieving input distributions for the MIMO-BF channel, see Proposition 22 for the details.

The following theorem summarizes our current knowledge of $v^*(n_t, T)$.

Theorem 5. *For any pair of positive integers n_t, T we have*

$$v^*(T, n_t) = v^*(n_t, T) \leq n_t T \min(n_t, T) . \tag{22}$$

If $n_t \leq \rho(T)$ or $T \leq \rho(n_t)$ then a full-rate orthogonal design is dispersion-optimal and

$$v^*(n_t, T) = n_t T \min(n_t, T) . \tag{23}$$

If instead $n_t > \rho(T)$ and $T > \rho(n_t)$ then for a jointly-Gaussian capacity-achieving input X we have[3]

$$\frac{n_t^2}{2P^2} \mathrm{Var}(\|X\|_F^2) < n_t T \min(n_t, T) . \tag{24}$$

Finally, if $n_t \leq T$ and (23) holds, then $v^*(n_t', T) = n_t'^2 T$ for any $n_t' \leq n_t$ (and similarly with the roles of n_t and T switched).

Note that the $\rho(n)$ function is monotonic in even values of n (and is 1 for n odd), and $\rho(n) \to \infty$ along even n. Therefore, for any number of transmit antennas n_t, there is a large enough T such that $n_t \leq \rho(T)$, in which case an $n_t \times T$ full rate orthogonal design achieves the optimal $v^*(n_t, T)$.

III. PRELIMINARY RESULTS

The section gives some results that will be useful for the achievability and converse proofs (Theorem 16 and Theorem 19, respectively), along with generally aiding our understanding of the MIMO-BF channel at finite blocklength. The results in this section and where they are used is summarized as follows:

- Theorem 6 gives a characterization of the caids for MIMO-BF channel. While all caids give the same capacity (by definition), when the channel matrix is rank 1, they do not all yield the same dispersion. This characterization is needed to reason about the minimizers in (8), especially in the rank 1 case.
- Proposition 8 computes variance $V_n(x^n)$ of information density conditioned on the channel input x^n. A key characteristic of the fading channel is that $V_n(x^n)$ varies as x^n moves around the input space, which does not happen in DMC's or the AWGN channel. This variation in $V_n(x^n)$ poses additional challenges in the converse proof, where we partition the codebook based on thresholding $V_n(x^n)$ (see the proof of Theorem 19 for details). Knowledge of $V_n(x^n)$ will also allow us to understand when the information density can be well approximated by a Gaussian (see Lemma 13).
- Propositions 11 and 12 explicitly give the expression for the dispersion found from the achievability and converse proofs for the $\mathrm{rank}(H) > 1$ and $\mathrm{rank}(H) \leq 1$ case, respectively. These expressions show how the dispersion depends on n_t, n_r, T, P, and are the contents of Theorems 2 and 3 above.

A. Known results: capacity and capacity achieving output distribution

First we review a few known results on the MIMO-BF channel. Since the channel is memoryless, the capacity is given by

$$C = \frac{1}{T} \max_{P_X : \mathbb{E}[\|X\|_F^2] \leq TP} I(X; Y, H) . \tag{25}$$

---

[3]So that in these cases the bound (22) is either non-tight, or is achieved by a non-jointly-Gaussian caid.

It was shown by Telatar [6] that whenever distribution of H is isotropic, the input $X \in \mathbb{R}^{n_t \times T}$ with entry i, j given by

$$X_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{P}{n_t}\right), \tag{26}$$

is a maximizer, resulting in the capacity formula (6). The distribution induced by a caid at the channel output (Y, H) is called the capacity achieving output distribution (caod). A classical fact is that, while there may be many caids, the caod is unique, e.g. [25, Section 4.4]. Thus, from (26) we infer that the caod is given by

$$P_{Y,H}^* \overset{\triangle}{=} P_H P_{Y|H}^*, \tag{27}$$

$$P_{Y|H}^* \overset{\triangle}{=} \prod_{j=1}^{T} P_{Y^{(j)}|H}^*, \tag{28}$$

$$P_{Y^{(j)}|H=h}^* \overset{\triangle}{=} \mathcal{N}\left(0, I_{n_r} + \frac{P}{n_t}hh^T\right), \tag{29}$$

$Y = [Y^{(1)}, \ldots, Y^{(T)}]$, where $Y^{(j)}$ is j-th column of Y, which, as we specified in (3), is a $n_r \times T$ matrix.

B. Capacity achieving input distributions

A key feature of the MIMO-BF channel is that it has many caids, whereas many commonly studied channels (e.g. BSC, BEC, AWGN) have a unique caid. Understanding the set of distributions that achieve capacity is essential for reasoning about the minimizer of the condition variance in (8). The following theorem characterizes the set of caids for the MIMO-BF channel. Somewhat surprisingly, for the case of rank-1 H (e.g. for MISO) there are multiple non-trivial jointly Gaussian caids with different correlation structures. For example, space-time block codes can achieve the capacity in the rank 1 case, but do not achieve capacity when the rank is 2 or greater e.g. [26].

Theorem 6.

1) *Every caid X satisfies*

$$\forall a \in \mathbb{R}^{n_t}, b \in \mathbb{R}^T: \qquad \sum_{i=1}^{n_t} \sum_{j=1}^{T} a_i b_j X_{i,j} \sim \mathcal{N}\left(0, \frac{P}{n_t}\|a\|_2^2 \|b\|_2^2\right). \tag{30}$$

If $\mathbb{P}[\operatorname{rank} H \leq 1] = 1$ then condition (30) is also sufficient for X to be caid.

2) *Let $X = \begin{pmatrix} R_1 \\ \cdots \\ R_{n_t} \end{pmatrix}$ be decomposed into rows R_i. If X is caid, then each $R_i \sim \mathcal{N}(0, \frac{P}{n_t}I_T)$ (i.i.d. Gaussian) and*

$$\mathbb{E}[R_i^T R_i] = \frac{P}{n_t}I_T, \qquad\qquad i = 1, \ldots, n_t \tag{31}$$

$$\mathbb{E}[R_i^T R_j] = -\mathbb{E}[R_j^T R_i], \qquad\qquad i \neq j. \tag{32}$$

If X is jointly zero-mean Gaussian and $\mathbb{P}[\operatorname{rank} H \leq 1] = 1$, then (31)-(32) are sufficient for X to be caid.

3) *Let $X = (C_1 \ldots C_T)$ be decomposed into columns C_j. If X is caid, then each $C_j \sim \mathcal{N}(0, \frac{P}{n_t}I_{n_t})$ (i.i.d. Gaussian) and*

$$\mathbb{E}[C_i C_i^T] = \frac{P}{n_t}I_{n_t}, \qquad\qquad i = 1, \ldots, T \tag{33}$$

$$\mathbb{E}[C_i C_j^T] = -\mathbb{E}[C_j C_i^T], \qquad\qquad i \neq j. \tag{34}$$

If X is jointly zero-mean Gaussian and $\mathbb{P}[\operatorname{rank} H \leq 1] = 1$, then (33)-(34) are sufficient for X to be caid.

4) *When $\mathbb{P}[\operatorname{rank} H > 1] > 0$, any caid has pairwise independent rows:*

$$R_i \perp\!\!\!\perp R_j \sim \mathcal{N}\left(0, \frac{P}{n_t}I_T\right) \qquad \forall i \neq j \tag{35}$$

and in particular

$$X_{i,j} \perp\!\!\!\perp X_{k,l} \qquad \forall (i,j) \neq (k,l). \tag{36}$$

Therefore, among jointly Gaussian X the i.i.d. $X_{i,j}$ is the unique caid.

5) *There exist non-Gaussian caids if and only if $\mathbb{P}[\operatorname{rank} H \geq \min(n_t, T)] = 0$.*

Remark 4. (Special case of rank-1 H) In the MISO case when $n_t > 1$ and $n_r = 1$ (or more generally, $\operatorname{rank} H \leq 1$ a.s.), there is not only a multitude of caids, but in fact they can have non-trivial correlations between entries of X (and this is ruled out by (36) for all other cases). As an example, for the $n_t = T = 2$ case, any of the following random matrix-inputs X (parameterized by $\rho \in [-1, 1]$) is a Gaussian caid:

$$X = \sqrt{\frac{P}{2}} \left[\begin{array}{cc} \xi_1 & -\rho\xi_2 + \sqrt{1-\rho^2}\xi_3 \\ \xi_2 & \rho\xi_1 + \sqrt{1-\rho^2}\xi_4 \end{array} \right], \tag{37}$$

where $\xi_1, \xi_2, \xi_3, \xi_4 \sim \mathcal{N}(0,1)$ i.i.d.. In particular, there are caids for which not all entries of X are pairwise independent.

Remark 5. Another way to state conditions (31)-(32) is: all elements in a row (resp. column) are pairwise independent $\sim \mathcal{N}(0, \frac{P}{n_t})$ and each 2×2 minor has antipodal correlation for the two diagonals. In particular, if X is caid, then X^T and any submatrix of X are caids too (for different n_t and T).

Proof. We will rely repeatedly on the following observations:

1) if A, B are two random vectors in $\mathbb{R}^n$ then and for any $v \in \mathbb{R}^n$ we have

$$\forall v \in \mathbb{R}^n : v^T A \stackrel{d}{=} v^T B \qquad \Longleftrightarrow \qquad A \stackrel{d}{=} B. \tag{38}$$

This is easy to show by computing characteristic functions.

2) If A, B are two random vectors in $\mathbb{R}^n$ independent of $Z \sim \mathcal{N}(0, I_n)$, then

$$A + Z \stackrel{d}{=} B + Z \qquad \Longleftrightarrow \qquad A \stackrel{d}{=} B. \tag{39}$$

This follows from the fact that the characteristic function of Z is nowhere zero.

3) For two matrices $Q_1, Q_2 \in \mathbb{R}^{n \times n}$ we have

$$\forall v \in \mathbb{R}^n : v^T Q_1 v = v^T Q_2 v \qquad \Longleftrightarrow \qquad Q_1 + Q_1^T = Q_2 + Q_2^T. \tag{40}$$

This follows from the fact that a quadratic form that is zero everywhere on $\mathbb{R}^n$ must have all coefficients equal to zero.

Part 1 (necessity). Recall that the caod is unique and given by (27). Thus an input X is caid iff for P_H-almost every $h_0 \in \mathbb{R}^{n_r \times n_t}$ we have

$$h_0 X + Z \stackrel{d}{=} h_0 G + Z, \tag{41}$$

where G is an $n_t \times T$ matrix with i.i.d. $\mathcal{N}(0, P/n_t)$ entries (for sufficiency, just write $I(X; Y, H) = h(Y|H) - h(Z)$ with $h(\cdot)$ denoting differential entropy). We will argue next that (41) implies (under isotropy assumption on P_H) that

$$\forall a \in \mathbb{R}^{n_t} : \quad a^T X \stackrel{d}{=} a^T G. \tag{42}$$

From (38), (42) is equivalent to $\sum_{i,j} a_i b_j X_{i,j} \stackrel{d}{=} \sum_{i,j} a_i b_j G_{i,j}$ for all $b \in \mathbb{R}^{n_t}$.

Let E_0 be a P_H-almost sure subset of $\mathbb{R}^{n_t \times n_r}$ for which (41) holds. Let $O(n) = \{U \in \mathbb{R}^{n \times n} : U^T U = UU^T = I_n\}$ denote the group of orthogonal matrices, with the topology inherited from $\mathbb{R}^{n \times n}$. Let $\{U_k\}$ and $\{V_k\}$ for $k \in \{1, 2, \ldots\}$ be countable dense subsets of $O(n_t)$ and $O(n_r)$, respectively. (These exist since $\mathbb{R}^{n^2}$ is a second-countable topological space). By isotropy of P_H we have $P_H[U_k(E_0)V_l] = 1$ and therefore

$$E \overset{\triangle}{=} E_0 \cap \bigcap_{k=1, l=1}^{\infty} U_k(E_0)V_l \tag{43}$$

is also almost sure: $P_H[E] = 1$. Here, $U_k(E_0)$ denotes the image of E_0 under U_k. By assumption (4), E must contain a non-zero element h_0, for otherwise we would have $P_H[0] = 1$, contradicting (4). Consequently, by isotropy $U_k h_0 V_l$ is also in E. From (39) and (41), we conclude that

$$U_k h_0 V_l X \overset{d}{=} U_k h_0 V_l G \qquad \forall k, l \,.$$

Arguing by continuity and using density of $\{U_k\}$ and $\{V_l\}$, this implies also

$$U h_0 V X \overset{d}{=} U h_0 V G \qquad \forall U \in O(n_t), V \in O(n_r) \,. \tag{44}$$

In particular, for any $a \in \mathbb{R}^{n_t}$ there must exist a choice of U, V such that $U h_0 V$ has the top row equal to $c_0 a^T$ for some constant $c_0 > 0$. Choosing these U, V in (44) and comparing distributions of top rows, we conclude (42) after scaling by $1/c_0$.

Part 1 (sufficiency). Suppose $\mathbb{P}[\text{rank } H \leq 1] = 1$. Then our goal is to show that (42) implies that X is a caid. To that end, it is sufficient to show $h_0 X \overset{d}{=} h_0 G$ for all rank-1 h_0. In the special case

$$h_0 = \begin{pmatrix} a^T \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

the claim follows directly from (42). Every other rank-1 h_0' can be decomposed as $h_0' = U h_0$ for some matrix U, and thus again we get $U h_0 X \overset{d}{=} U h_0 G$, concluding the proof.

Parts 2 and 3 (necessity). From part 1 we have that for every a, b we must have $a^T X b \sim \mathcal{N}(0, \|a\|_2^2 \|b\|_2^2 \frac{P}{n_t})$. Computing expected square we get

$$\mathbb{E}\left[(a^T X b)^2\right] = \frac{P}{n_t} \left(\sum_i a_i^2\right) \left(\sum_j b_j^2\right) \,. \tag{45}$$

Thus, expressing the left-hand side in terms of rows R_i as $a^T X = \sum_i a_i R_i$ we get

$$b^T \left\{ \mathbb{E}\left[\left(\sum_i a_i R_i\right)^T \left(\sum_i a_i R_i\right) \right] \right\} b = b^T \left(\sum_i a_i^2 I_T\right) b \,,$$

and thus by (40) we conclude that for all a:

$$\mathbb{E}\left[\left(\sum_i a_i R_i\right)^T \left(\sum_i a_i R_i\right) \right] = \left(\sum_i a_i^2\right) I_T \,.$$

Each entry of the $T \times T$ matrices above is a quadratic form in a and thus again by (40) we conclude (31)-(32). Part 3 is argued similarly with roles of a and b interchanged.

Parts 2 and 3 (sufficiency). When H is (at most) rank-1, we have from part 1 that it is sufficient to show that $a^T X b \sim \mathcal{N}(0, \|a\|_2^2 \|b\|_2^2 \frac{P}{n_t})$. When X is jointly zero-mean Gaussian, we have $a^T X b$ is zero-mean

Gaussian and so we only need to check its second moment satisfies (45). But as we just argued, (45) is equivalent to either (31)-(32) or (33)-(34).

Part 4. As in Part 1, there must exist $h_0 \in \mathbb{R}^{n_r \times n_t}$ such that (44) holds and $\operatorname{rank} h_0 > 1$. Thus, by choosing U, V we can diagonalize h_0 and thus we conclude any pair of rows R_i, R_j must be independent.

Part 5. This part is never used in subsequent parts of the paper, so we only sketch the argument and move the most technical part of the proof to Appendix A. Let $\ell = \max\{r : \mathbb{P}[\operatorname{rank} H \geq r] > 0\}$. Then arguing as for (44) we conclude that X is caid if and only if for any h with $\operatorname{rank} h \leq \ell$ we have

$$hX \stackrel{d}{=} hG .$$

In other words, we have

$$\sum_{i,j} a_{i,j} X_{i,j} \stackrel{d}{=} \sum_{i,j} G_{i,j} \qquad \forall a \in \mathbb{R}^{n_t \times T} : \operatorname{rank} a \leq \ell . \tag{46}$$

If $\ell = \min(n_t, T)$, then rank condition on a is not active and hence, we conclude by (38) that $X \stackrel{d}{=} G$. So assume $\ell < \min(n_t, T)$. Note that (46) is equivalent to the condition on characteristic function of X as follows:

$$\mathbb{E}\left[e^{i \sum_{i,j} a_{i,j} X_{i,j}}\right] = e^{-\frac{P}{2n_t} \sum_{i,j} a_{i,j}^2} \qquad \forall a : \operatorname{rank} a \leq \ell . \tag{47}$$

It is easy to find polynomial (in $a_{i,j}$) that vanishes on all matrices of rank $\leq \ell$ (e.g. take the product of all $\ell \times \ell$ minors). Then Proposition 24 in Appendix A constructs non-Gaussian X satisfying (47) and hence (46). $\qquad \square$

C. Information density and its moments

In finite blocklength analysis, a key object of study is the information density, along with its first and second moments. In this section we'll find expressions for these moments, along with showing when the information density is asymptotically normal.

It will be convenient to assume that the matrix H is represented as

$$H = U\Lambda V^T , \tag{48}$$

where U, V are uniformly distributed on $O(n_r)$ and $O(n_t)$ (which follows from the isotropic assumption H), respectively,[4] and Λ is the $n_r \times n_t$ diagonal matrix with diagonal entries $\{\Lambda_i, i = 1, \ldots, n_{\min}\}$. Joint distribution of $\{\Lambda_i\}$ depends on the fading model. It does not matter for our analysis whether Λ_i's are sorted in some way, or permutation-invariant.

For the MIMO-BF channel, let P_{YH}^* denote the caod (27). To compute the information density with respect to P_{YH}^* (for a single T-block of symbols) as defined in (9), denote $y = hx + z$ and write an SVD decomposition for matrix h as

$$h = u\lambda v^T ,$$

where $u \in O(n_r)$, $v \in O(n_t)$ and λ is an $n_r \times n_t$ matrix which is zero except for the diagonal entries, which are equal to $\lambda_1, \ldots, \lambda_{n_{\min}}$. Note that this representation is unique up to permutation of $\{\lambda_j\}$, but the choice of this permutation will not affect any of the expressions below. With this decomposition we have:

$$i(x; y, h) \stackrel{\triangle}{=} \frac{T}{2} \log \det \left(I_{n_r} + \frac{P}{n_t} hh^T\right)$$
$$+ \frac{\log e}{2} \sum_{j=1}^{n_{\min}} \frac{\lambda_j^2 \|v_j^T x\|^2 + 2\lambda_j \langle v_j^T x, \tilde{z}_j \rangle - \frac{P}{n_t} \lambda_j^2 \|\tilde{z}_j\|^2}{1 + \frac{P}{n_t} \lambda_j^2} \tag{49}$$

---

[4]Recall that $O(m) = \{A \in \mathbb{R}^{m \times m} : AA^T = A^T A = I_m\}$ is the space of all orthogonal matrices. This space is compact in a natural topology and admits a Haar probability measure.

where we denoted by v_j the j-th column of V, and have set $\tilde{z} = u^T z$, with $\tilde{z}_j$ representing the j-th row of $\tilde{z}$. The definition naturally extends to blocks of length nT additively:

$$i(x^n; y^n, h^n) \triangleq \sum_{j=1}^{n} i(x_j; y_j, h_j).$$

(50)

We compute the (conditional) mean of information density to get

$$D_n(x^n) \triangleq \frac{1}{nT} \mathbb{E}\left[i(X^n; Y^n, H^n)|X^n = x^n\right]$$

(51)

$$= C(P) + \frac{\sqrt{\frac{\eta_2}{2}}}{n_t nT} \sum_{j=1}^{n} (\|x_j\|_F^2 - TP),$$

(52)

where we used the following simple fact:

Lemma 7. *Let $U \in \mathbb{R}^{1 \times n_t}$ be uniformly distributed on the unit sphere, and $x \in \mathbb{R}^{n_t \times T}$ be a fixed matrix, then*

$$\mathbb{E}[\|Ux\|^2] = \frac{\|x\|_F^2}{n_t}$$

(53)

Proof. Note that by additivity of $\|Ux\|^2$ across columns, it is sufficient to consider the case $T = 1$, for which the statement is clear from symmetry. $\qquad \square$

Remark 6. A simple consequence of Lemma 7 is $\mathbb{E}[\|Hx\|_F^2] = \mathbb{E}[\|H\|_F^2]\frac{\|x\|_F^2}{n_t}$, which follows from considering the SVD of H.

Proposition 8. *Let $V_n(x^n) \triangleq \frac{1}{nT}\mathrm{Var}(i(X^n; Y^n, H^n)|X^n = x^n)$, then we have*

$$V_n(x^n) = \frac{1}{n} \sum_{j=1}^{n} V_1(x_j),$$

(54)

where the function $V_1 : \mathbb{R}^{n_t \times T} \mapsto \mathbb{R}$ defined as $V_1(x) \triangleq \frac{1}{T}\mathrm{Var}(i(X; Y, H)|X = x)$ is given by

$$V_1(x) = T\mathrm{Var}\left(C_r(H, P)\right)$$

(55)

$$+ \sum_{i=1}^{n_{\min}} \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\Lambda_i^2\right)\right]$$

(56)

$$+ \eta_5 \left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)$$

(57)

$$+ \eta_3 \left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)^2$$

(58)

$$+ \eta_4 \left(\|xx^T\|_F^2 - \frac{1}{n_t}\|x\|_F^4\right)$$

(59)

11

where $c(\cdot)$ was defined in (13) *and*

$$C_r(H, P) \triangleq \frac{1}{2} \log \det \left(I_{n_r} + \frac{P}{n_t} H H^T \right) = \sum_{i=1}^{n_{\min}} C_{AWGN} \left(\frac{P}{n_t} \Lambda_i^2 \right) \tag{60}$$

$$\eta_3 \triangleq \frac{\log^2 e}{4} \mathrm{Var} \left(\sum_{k=1}^{n_{\min}} c(\Lambda_k^2) \right) \tag{61}$$

$$\eta_4 \triangleq \frac{\log^2 e}{2 n_t (n_t + 2)} \left(\mathbb{E} \left[\sum_{i=1}^{n_{\min}} c^2(\Lambda_i^2) \right] - \frac{1}{(n_t - 1)} \sum_{i \neq j} \mathbb{E} \left[c(\Lambda_i^2) c(\Lambda_j^2) \right] \right) \tag{62}$$

$$\eta_5 \triangleq \frac{\log e}{2} \mathrm{Cov} \left(C_r(H, P), \sum_{k=1}^{n_{\min}} c(\Lambda_k^2) \right) + \frac{\log^2 e}{T} \sum_{k=1}^{n_{\min}} \mathbb{E} \left[\frac{\Lambda_k^2}{\left(1 + \frac{P}{n_t} \Lambda_k^2 \right)^2} \right] . \tag{63}$$

Remark 7. Every term in the definition of $V_1(x)$ (except the one with η_5) is non-negative (for η_4-term, see (88)). The η_5-term will not be important because for inputs satisfying power-constraint with equality it vanishes. Note also that the first term in (63) can alternatively be given as

$$\mathrm{Cov} \left(C_r(H, P), \sum_{k=1}^{n_{\min}} c(\Lambda_k^2) \right) = n_t \frac{d}{dP} \mathrm{Var} \left[C_r(H, P) \right] .$$

Proof. From (49), we have the form of the information density. First note that the information density over n channel uses decomposes into a sum of n independent terms,

$$i(x^n; Y^n, H^n) = \sum_{j=1}^{n} i(x_j, Y_j, H_j) . \tag{64}$$

As such, the variance conditioned on x^n also decomposes as

$$\mathrm{Var}(i(x^n; Y^n, H^n)) = \sum_{j=1}^{n} \mathrm{Var}(i(x_j; Y_j, H_j)) , \tag{65}$$

from which (54) follows. Because the variance decomposes as a sum in (65), we focus on only computing $\mathrm{Var}(i(x; Y, H))$ for a single coherent block. Define

$$f(h) \triangleq T C_r(h, P) \tag{66}$$

$$g(x, h, z) \triangleq \frac{\log e}{2} \sum_{k=1}^{n_{\min}} \frac{\Lambda_k^2 \| v_k^T x \|^2 + 2 \Lambda_k \langle v_k^T x, \tilde{z}_k \rangle - \frac{P}{n_t} \Lambda_k^2 \| \tilde{z}_k \|^2}{1 + \frac{P}{n_t} \Lambda_k^2} \tag{67}$$

so that $i(x; y, h) = f(h) + g(x, h, z)$ in notation from (49). With this, the quantity of interest is

$$\mathrm{Var}(i(x, Y, H)) = \mathrm{Var}(f(H)) + \mathrm{Var}(g(x, H, Z)) + \mathrm{Cov}(f(H), g(x, H, Z)) \tag{68}$$

$$= \underbrace{\mathrm{Cov}(f(H), g(x, H, Z))}_{\triangleq T_1} + \underbrace{\mathrm{Var}(f(H))}_{\triangleq T_2} + \underbrace{\mathrm{Var}\mathbb{E}[g(x, HZ)|H]}_{\triangleq T_3} + \underbrace{\mathbb{E}\mathrm{Var}(g(x, H, Z)|H)}_{\triangleq T_4} \tag{69}$$

where (69) follows from the identity

$$\mathrm{Var}(g(x, H, Z)) = \mathbb{E}\mathrm{Var}(g(x, H, Z)|H) + \mathrm{Var}\mathbb{E}[g(x, H, Z)|H] . \tag{70}$$

Below we show that T_1 and T_3 corresponds to (57), T_2 corresponds to (55), T_4 corresponds to (56), and T_3 corresponds to (58) and (59). We evaluate each term separately.

$$T_1 = \text{Cov}(f(H), g(x, H, Z)) \tag{71}$$

$$= \mathbb{E}\left[(f(H) - \mathbb{E}[f(H)])(g(x, H, Z) - \mathbb{E}[g(x, H, Z)])\right] \tag{72}$$

$$= \frac{\log e}{2}\left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)\sum_{k=1}^{n_{\min}} \mathbb{E}\left[(f(H) - \mathbb{E}[f(H)])(c(\Lambda_k^2) - \mathbb{E}[c(\Lambda_k^2)])\right] \tag{73}$$

$$= \frac{\log e}{2}\left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)\sum_{k=1}^{n_{\min}} \text{Cov}\left(f(H), c(\Lambda_k^2)\right) \tag{74}$$

where (73) follows from noting that

$$\mathbb{E}\left[g(x, H, Z)|H\right] = \sum_{k=1}^{n_{\min}}\left(\|V_k^T x\|^2 - \frac{TP}{n_t}\right)c(\Lambda_k^2)\frac{\log e}{2}. \tag{75}$$

Now, since V_k is independent from Λ_k by the rotational invariance assumption, we have that $f(H)$ is independent from V_k, since $f(H)$ only depends on H through its eigenvalues, see (60). We are only concerned with the expectation over $g(x, H, Z)$ in (72), which reduces to

$$\mathbb{E}\left[g(x, H, Z) - \mathbb{E}[g(x, H, Z)]|\Lambda_1, \ldots, \Lambda_{n_{\min}}\right] = \left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)\sum_{k=1}^{n_{\min}} c(\Lambda_k^2) - \mathbb{E}[c(\Lambda_k)^2]\frac{\log e}{2}, \tag{76}$$

giving (73).

Next, T_2 in (69) becomes

$$T_2 = \text{Var}(f(H)) \tag{77}$$

$$= T^2\text{Var}\left(\sum_{k=1}^{n_{\min}} C_{AWGN}\left(\frac{P}{n_t}\Lambda_k^2\right)\right). \tag{78}$$

For T_3 in (69), we obtain

$$T_3 = \mathbb{E}\text{Var}(g(x, H, Z)|H) \tag{79}$$

$$= \frac{\log^2 e}{4}\mathbb{E}\left[\sum_{k=1}^{n_{\min}} \frac{4\Lambda_k^2\|V_k^T x\|^2 + 2T\left(\frac{P}{n_t}\right)^2\Lambda_k^4}{\left(1 + \frac{P}{n_t}\Lambda_k\right)^2}\right] \tag{80}$$

$$= \frac{\log^2 e}{2}\sum_{k=1}^{n_{\min}} T\mathbb{E}\left[\frac{2\frac{TP}{n_t}\Lambda_k^2 + T\left(\frac{P}{n_t}\right)^2\Lambda_k^4}{\left(1 + \frac{P}{n_t}\Lambda_k\right)^2}\right] + 2\mathbb{E}\left[\frac{\frac{\|x\|_F^2}{n_t}\Lambda_k^2 - \frac{TP}{n_t}\Lambda_k^2}{\left(1 + \frac{P}{n_t}\Lambda_k^2\right)^2}\right] \tag{81}$$

$$= T\sum_{k=1}^{n_{\min}} V_{AWGN}\left(\frac{P}{n_t}\Lambda_k^2\right) + \log^2(e)\left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)\mathbb{E}\left[\frac{\Lambda_k^2}{\left(1 + \frac{P}{n_t}\Lambda_k^2\right)^2}\right] \tag{82}$$

where
- (80) follows from taking the variance over $\tilde{Z}$ (recall $\tilde{Z} = U^T Z$ in (49)).
- (81) follows from Lemma 7 applied to $\mathbb{E}[\|V_k^T x\|^2]$, and adding and subtracting the term

$$\log^2(e)\mathbb{E}\left[\frac{\frac{TP}{n_t}\Lambda_k^2}{\left(1 + \frac{P}{n_t}\Lambda_k^2\right)^2}\right]. \tag{83}$$

Continuing with T_3 from (69),

$$T_3 = \text{Var}\mathbb{E}[g(x, H, Z)|H] \tag{84}$$

$$= \text{Var}\left(\frac{\log e}{2} \sum_{k=1}^{n_{\min}} c(\Lambda_k^2)\left(\|V_k^T x\|^2 - \frac{TP}{n_t}\right)\right) \tag{85}$$

$$= \eta_3 \left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)^2$$

$$+ \frac{\log^2 e}{4}\mathbb{E}\text{Var}\left(\sum_{k=1}^{n_{\min}} c(\Lambda_k^2)\|V_k^T x\|^2 \middle| \Lambda_1, \ldots, \Lambda_{n_{\min}}\right) \tag{86}$$

where
- (85) follows from taking the expectation over $\tilde{Z}$,
- (86) follows from applying the variance identity (70) with respect to V and $\Lambda_1, \ldots, \Lambda_{n_{\min}}$, as well as recalling (61).

We are left to show that the term (86) equals (59). To that end, define

$$\phi(x) \triangleq \mathbb{E}\text{Var}\left(\sum_{k=1}^{n_{\min}} c(\Lambda_k^2)\|V_k^T x\|^2 \middle| \Lambda_1, \ldots, \Lambda_{n_{\min}}\right) \tag{87}$$

$$= \sum_{k=1}^{n_{\min}} \mathbb{E}[c^2(\Lambda_k^2)]\text{Var}\left(\|V_k^T x\|^2\right) + \sum_{k \neq l}^{n_{\min}} \mathbb{E}[c(\Lambda_k^2)c(\Lambda_l^2)]\text{Cov}(\|V_k^T x\|^2, \|V_l^T x\|^2) . \tag{88}$$

We will finish the proof by showing

$$\phi(x) = \frac{4}{\log^2 e}\eta_4\left(\|xx^T\|_F^2 - \frac{1}{n_t}\|x\|_F^4\right) .$$

To that end, we first compute moments of V drawn from the Haar measure on the orthogonal group.

Lemma 9. *Let V be drawn from the Haar measure on $O(n)$, then for $i, j, k, l = 1, \ldots, n$ all unique,*

$$\mathbb{E}[V_{ij}^2] = \frac{1}{n} \tag{89}$$

$$\mathbb{E}[V_{ij}V_{ik}] = 0 \tag{90}$$

$$\mathbb{E}[V_{ij}^2 V_{ik}^2] = \frac{1}{n(n+2)} \tag{91}$$

$$\mathbb{E}[V_{ij}^2 V_{kl}^2] = \frac{n+1}{n(n-1)(n+2)} \tag{92}$$

$$\mathbb{E}[V_{ij}^4] = \frac{3}{n(n+2)} \tag{93}$$

$$\mathbb{E}[V_{ij}V_{ik}V_{lj}V_{lk}] = \frac{-1}{n(n-1)(n+2)} . \tag{94}$$

Proof of this Lemma is given below.

First, note that the variance $\mathrm{Var}(\|V_k^T x\|^2)$ does not depend on k, since the marginal distribution of each V_k is uniform on the unit sphere. Hence below we only consider V_1. We obtain

$$\mathrm{Var}(\|V_1^T x\|^2) = \mathbb{E}[\|V_1^T x\|^4] - \mathbb{E}^2[\|V_1^T x\|^2] \tag{95}$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{T}\sum_{j=1}^{n_t}\sum_{k=1}^{n_t} V_{j1}V_{k1}x_{ji}x_{ki}\right)^2\right] - \frac{\|x\|_F^4}{n_t^2} \tag{96}$$

$$= \mathbb{E}\left[\sum_{j=1}^{n_t}\sum_{k=1}^{n_t}\sum_{l=1}^{n_t}\sum_{m=1}^{n_t} V_{j1}V_{k1}V_{l1}V_{m1}\langle r_j, r_k\rangle\langle r_l, r_m\rangle\right] \tag{97}$$

where r_j denotes the j-th row of x. Now it is a matter counting similar terms.

$$\mathbb{E}[\|V_1^T x\|^4] = \sum_{j=1}^{n_t}\mathbb{E}[V_{j1}^4]\|r_j\|^4 + 2\sum_{j\neq k}^{n_t}\mathbb{E}[V_{j1}^2 V_{k1}^2]\langle r_j, r_k\rangle^2 + \sum_{j\neq k}^{n_t}\mathbb{E}[V_{j1}^2 V_{k1}^2]\|r_j\|^2\|r_k\|^2 \tag{98}$$

$$= \frac{3}{n_t(n_t+2)}\sum_{j=1}^{n_t}\|r_j\|^4 + \frac{2}{n_t(n_t+2)}\sum_{j\neq k}^{n_t}\langle r_j, r_k\rangle^2 + \frac{1}{n_t(n_t+2)}\sum_{j\neq k}\|r_j\|^2\|r_k\|^2 \tag{99}$$

$$= \frac{1}{n_t(n_t+2)}\left(\|x\|_F^4 + 2\|xx^T\|_F^2\right) \tag{100}$$

where

- (98) follows from collecting like terms from the summation in (97).
- (99) uses Lemma 9 to compute each expectation.
- (100) follows from realizing that

$$\|x\|_F^4 = \left(\sum_{j=1}^{n_t}\|r_j\|^2\right)^2 = \sum_{j=1}^{n_t}\|r_j\|^4 + \sum_{j\neq k}^{n_t}\|r_j\|^2\|r_k\|^2 \tag{101}$$

$$\|xx^T\|_F^2 = \sum_{j=1}^{n_t}\sum_{k=1}^{n_t}\langle r_j, r_k\rangle^2 = \sum_{j=1}^{n_t}\|r_j\|^4 + \sum_{j\neq k}^{n_t}\langle r_j, r_k\rangle^2 \tag{102}$$

Plugging this back into (95) yields the variance term,

$$\mathrm{Var}(\|V_1^T x\|^2) = \frac{1}{n_t(n_t+2)}\left(\|x\|_F^4 + 2\|xx^T\|_F^2\right) - \frac{\|x\|_F^4}{n_t^2} = \frac{2}{n_t(n_t+2)}\left(\|xx^T\|_F^2 - \frac{\|x\|_F^4}{n_t}\right). \tag{103}$$

Now we compute the covariance term from (88) in a similar way. By symmetry of the columns of V, we can consider only the covariance between $\|V_1^T x\|^2$ and $\|V_2^T x\|^2$, i.e.

$$\mathrm{Cov}(\|V_1^T x\|^2, \|V_2^T x\|^2) = \mathbb{E}[\|V_1^2 x\|^2\|V_2^T x\|^2] - \frac{\|x\|_F^4}{n_t^2}. \tag{104}$$

Expanding the expectation, we get

$$\mathbb{E}[\|V_1^T x\|^2\|V_2^T x\|^2] \tag{105}$$

$$= \sum_{j,k,l,m}\mathbb{E}[V_{1j}V_{1k}V_{2l}V_{2m}]\langle r_j, r_k\rangle\langle r_l, r_m\rangle \tag{106}$$

$$= \sum_{j=1}^{n_t}\mathbb{E}[V_{1j}^4]\|r_j\|^4 + \sum_{j\neq k}\mathbb{E}[V_{1j}^2 V_{2k}^2]\|r_j\|^2\|r_k\|^2 + 2\sum_{j\neq k}\mathbb{E}[V_{1j}V_{1k}V_{2j}V_{2k}]\langle r_j, r_k\rangle^2 \tag{107}$$

$$= \frac{1}{n_t(n_t+2)}\sum_{j=1}^{n_t}\|r_j\|^4 + \frac{n_t+1}{(n_t-1)n_t(n_t+2)}\sum_{j\neq k}\|r_j\|^2\|r_k\|^2 - \frac{2}{(n_t-1)n_t(n_t+2)}\sum_{j\neq k}\langle r_j, r_k\rangle^2 \tag{108}$$

$$= \frac{1}{(n_t-1)n_t(n_t+2)}\left((n_t+1)\|x\|_F^4 - 2\|xx^T\|_F^2\right). \tag{109}$$

15

With this, we obtain from (104),

$$\text{Cov}(\|V_1^T x\|^2, \|V_2^T x\|^2) = \frac{2}{(n_t - 1)n_t(n_t + 2)} \left(\frac{\|x\|_F^4}{n_t} - \|xx^T\|_F^2 \right) \tag{110}$$

where the steps follow just as in the variance computation (98)-(100).

Finally, returning to (88), using the variance (103) and covariance (110), we obtain

$$\phi(x) = \frac{2}{n_t(n_t + 2)} \left(\|xx^T\|_F^2 - \frac{\|x\|_F^4}{n_t} \right) \left(\sum_{k=1}^{n_t} \mathbb{E}[c^2(\Lambda_k^2)] - \frac{1}{n_t - 1} \sum_{k \neq l} \mathbb{E}[c(\Lambda_k^2)c(\Lambda_l^2)] \right) . \tag{111}$$

Plugging this into (86) finishes the proof. $\qquad\square$

Proof of Lemma 9. We first note that all entries of V have identical distribution, since permutations of rows and columns leave the distribution invariant. Because of this, we can WLOG only consider $V_{11}, V_{12}, V_{21}, V_{22}$ to prove the lemma.

- (89) follows immediately from $\sum_{i=1}^n V_{ij}^2 = 1$ a.s.
- Let V_i, V_j be any two distinct columns of V, then (90) follows from

$$0 = \mathbb{E}[\langle V_i, V_j \rangle] = n\mathbb{E}[V_{11}V_{21}] \tag{112}$$

- For (91) and (94), let $E_1 = \mathbb{E}[V_{11}^4]$ and $E_2 = \mathbb{E}[V_{11}^2 V_{21}^2]$. The following relations between E_1, E_2 hold,

$$1 = \mathbb{E}\left[\left(\sum_{j=1}^n V_{1j}^2 \right)^2 \right] \tag{113}$$

$$= nE_1 + n(n-1)E_2 \tag{114}$$

and, noticing that multiplication of V by the matrix

$$\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & I_{n-2} \end{bmatrix} \tag{115}$$

where I_n is the $n \times n$ identity matrix. This is an orthogonal matrix, so we obtain the relation

$$E_1 = \mathbb{E}\left[\left(\frac{V_{11}}{\sqrt{2}} + \frac{V_{12}}{\sqrt{2}} \right)^4 \right] \tag{116}$$

$$= \frac{1}{2}E_1 + \frac{3}{2}E_2 \tag{117}$$

from which we obtain $E_1 = 3E_2$. With this and (114), we obtain

$$E_1 = \frac{3}{n(n+2)} \tag{118}$$

$$E_2 = \frac{1}{n(n+2)} \tag{119}$$

- For (92), take

$$E_3 = \mathbb{E}[V_{11}^2 V_{22}^2] \tag{120}$$

$$= \mathbb{E}\left[V_{11}^2 \left(1 - \sum_{j \neq 2}^n V_{2j}^2 \right) \right] \tag{121}$$

$$= \frac{1}{n} - \frac{1}{n(n+2)} - (n-2)E_3 . \tag{122}$$

Solving for E_3 yields (92).

- For (94), let V_1, V_2 denote the first and second column of V respectively, and let $E_4 = \mathbb{E}[V_{11}V_{12}V_{21}V_{22}]$, then (94) follows from

$$0 = \mathbb{E}[\langle V_1, V_2 \rangle^2] \tag{123}$$

$$= nE_2 + n(n-1)E_4 \ . \tag{124}$$

Using E_2 from (119) and solving for E_4 gives (94).

$\square$

The following propsition gives the value of the conditional variance of the information density when input distribution has i.i.d. $\mathcal{N}(0, P/n_t)$ entries. This will turn out to be the operational dispersion in the case where $\operatorname{rank} H > 1$.

Proposition 10. *Let $X^n = (X_1, \ldots, X_n)$ be i.i.d. with Telatar distribution* (26) *for each entry. Then*

$$\mathbb{E}\operatorname{Var}(i(X^n; Y^n, H^n)|X^n) = nTV_{iid}(P) , \tag{125}$$

where $V_{iid}(P)$ is the right-hand side of (12).

Proof. To show this, we take the expectation of the expression given in Proposition 8 when X^n has i.i.d. $\mathcal{N}(0, P/n_t)$ entries. The terms (55) and (56) do not depend on X^n, and these give us the first two terms in (12). (57) vanishes immediately, since $\mathbb{E}[\|X\|_F^2] = TP$ by the power constraint. It is left to compute the expectation over (58) and (59) from the expression in Proposition 8. Using identities for χ^2 distributed random variables (namely, $\mathbb{E}[\chi_k^2] = k$, $\operatorname{Var}(\chi_k^2) = 2k$), we get:

$$\frac{\eta_3}{n_t^2}\operatorname{Var}(\|X_1\|_F^2) = \frac{\eta_3}{n_t}\left(\frac{P}{n_t}\right)^2 2T \tag{126}$$

$$\mathbb{E}[\|X_1\|_F^4] = TP^2\left(T + \frac{2}{n_t}\right) \tag{127}$$

$$\mathbb{E}[\|X_1 X_1^T\|_F^2] = n_t T\left(\frac{P}{n_t}\right)^2 (1 + T + n_t) \tag{128}$$

$$\mathbb{E}\left[\|X_1 X_1^T\|_F^2 - \frac{\|X_1\|_F^4}{n_t}\right] = T\left(\frac{P}{n_t}\right)^2 (n_t - 1)(n_t + 2) . \tag{129}$$

Hence, the sum of terms in (58) + (59) after taking expectation over X^n yields

$$T\left(\frac{P}{n_t}\right)^2 \left[2\frac{\eta_3}{n_t} + (n_t - 1)(n_t + 2)\eta_4\right] .$$

Introducing random variables $U_i = c(\Lambda_i^2)$ the expression in the square brackets equals

$$\frac{\log^2 e}{2}\frac{1}{n_t}\left[\operatorname{Var}\left(\sum_i U_i\right) + (n_t - 1)\sum_i \mathbb{E}[U_i^2] - \sum_{i \neq j}\mathbb{E}[U_i U_j]\right] . \tag{130}$$

At the same time, the third term in expression (12) is

$$\frac{\log^2 e}{2}\frac{1}{n_t}\left[n_t\sum_i \mathbb{E}[U_i^2] - \left(\sum_i \mathbb{E}[U_i]\right)^2\right] . \tag{131}$$

One easily checks that (130) and (131) are equal.

$\square$

The next proposition shows that, when the rank of H is larger than 1, the conditional variance in (8) is constant over the set of caids. Thus we can compute the conditional variance for the i.i.d. $\mathcal{N}(0, P/n_t)$ caid, and conclude that this expression is the minimizer in (8).

Proposition 11. *If $\mathbb{P}[\operatorname{rank} H > 1] > 0$, then for* <u>any</u> *caid $X \sim P_X$ we have*

$$\operatorname{Var}(X;Y,H)) = T\mathbb{E}\left[V_1(X)\right] = TV_{iid}(P).$$

In particular, the $V(P)$ defined as infimum over all caids (8) *satisfies $V(P) = V_{iid}(P)$.*

Proof. For any caid the term (57) vanishes. Let X^* be Telatar distributed. To analyze (58) we recall that from (36) we have

$$\mathbb{E}\left[\|X\|_F^4\right] = \sum_{i,j,i',j'} \mathbb{E}\left[X_{i,j}^2 X_{i',j'}^2\right] = \mathbb{E}\left[\|X^*\|_F^4\right].$$

For the term (59) we notice that

$$\|XX^T\|_F^2 = \sum_{i,j}\langle R_i, R_j\rangle^2,$$

where R_i is the i-th row of X. By (35) from Theorem 6 we then also have

$$\mathbb{E}\left[\|XX^T\|_F^2\right] = \mathbb{E}\left[\|X^*X^{*T}\|_F^2\right].$$

To conclude, $\mathbb{E}\left[V_1(X)\right] = \mathbb{E}\left[V_1(X^*)\right] = V_{iid}(P)$. $\qquad\qquad\square$

In the case where $\operatorname{rank} H \leq 1$, it turns out that the conditional variance *does* vary over the set of caids. The following proposition gives the expression for the conditional variance in this case, as a function of the caid.

Proposition 12. *If $\mathbb{P}[\operatorname{rank}(H) \leq 1] = 1$, then for any capacity achieving input X we have*

$$\frac{1}{T}\mathbb{E}\operatorname{Var}(i(X;Y,H)|X) = T\operatorname{Var}\left(C_{AWGN}\left(\frac{P}{n_t}\Lambda_1^2\right)\right) + \mathbb{E}V_{AWGN}\left(\frac{P}{n_t}\Lambda_1^2\right) \tag{132}$$

$$+ \eta_1\left(\frac{P}{n_t}\right)^2 - \frac{\eta_2}{2n_t^2 T}\operatorname{Var}(\|X\|_F^2) \tag{133}$$

where η_1, η_2 are defined in (14)-(15).

Proof. As in Prop. 10 we need to evaluate the expectation of terms in (57)-(59). Any caid X should satisfy $\mathbb{E}\left[\|X\|_F^2\right] = TP$ and thus the term (57) is zero. The term (58) can be expressed in terms of $\operatorname{Var}(\|X\|_F^2)$, but the (59) presents a non-trivial complication due to the presence of $\|XX^T\|_F^2$, whose expectation is possible (but rather tedious) to compute by invoking properties of caids established in Theorem 6. Instead, we recall that the sum (58)+(59) equals (86). Evaluation of the latter can be simplified in this case due to constraint on the rank of H. Overall, we get

$$\mathbb{E}\operatorname{Var}(i(X;Y,H)|X) = T^2\operatorname{Var}\left(C_{AWGN}\left(\frac{P}{n_t}\Lambda_1^2\right)\right) + T\mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\Lambda_1^2\right)\right] \tag{134}$$

$$+ \frac{\log^2 e}{4}\mathbb{E}\operatorname{Var}\left[c(\Lambda_1^2)\left(\|V_1^T X\|^2 - \frac{TP}{n_t}\right)\Big|X\right], \tag{135}$$

where $c(\cdot)$ is from (13). The last term in (135) can be written as

$$\mathbb{E}\left[c(\Lambda_1^2)^2\right]\mathbb{E}\left[\left(\|V_1^T X\|^2 - \frac{TP}{n_t}\right)^2\right] - \mathbb{E}^2[c(\Lambda_1^2)]\mathbb{E}\left[\left(\mathbb{E}[\|V_1^T X\|_F^2|X] - \frac{TP}{n_t}\right)^2\right] \tag{136}$$

which follows from the identity $\operatorname{Var}(AB) = \mathbb{E}[A^2]\mathbb{E}[B^2] - \mathbb{E}^2[A]\mathbb{E}^2[B]$ for independent A, B. The second term in (136) is easily handled since from Lemma 7 we have $\mathbb{E}[\|V_1^T X\|_F^2|X] = \|X\|_F^2/n_t$. To compute the first term in (136) recall from Theorem 6 that for any fixed unit-norm v and caid X we must have $v^T X \sim \mathcal{N}(0, P/n_t I_T)$. Therefore, we have

$$\mathbb{E}\left[\left(\|V_1^T X\|^2 - \frac{TP}{n_t}\right)^2\Big|V_1\right] = \frac{2TP^2}{n_t^2}.$$

Putting everything together we get that (136) equals

$$\mathbb{E}[c(\Lambda_1^2)^2] 2T \left(\frac{P}{n_t}\right)^2 - \mathbb{E}[c(\Lambda_1^2)]^2 \frac{1}{n_t^2} \text{Var}(\|X\|_F^2) \tag{137}$$

concluding the proof. □

The question at hand is: which input distribution X that achieves capacity minimizes (132)? Proposition 12 reduces this problem to maximizing $\text{Var}(\|X\|_F^2)$ over the set of capacity achieving input distributions. This will be analyzed in Section VI.

Finally, the following lemma computes the Berry Esseen constant. This is a technical result that will be needed for both the achievability and converse proofs.

Lemma 13. *Fix $x_1, \ldots, x_n \in \mathbb{R}^{n_t \times T}$ and let $W_j = i(x_j; Y_j, H_j)$, where Y_j, H_j are distributed as the output of channel* (3) *with input x_j. Define the Berry-Esseen ratio*

$$B_n(x^n) \triangleq \sqrt{n} \frac{\sum_{j=1}^n \mathbb{E}\left[|W_j - \mathbb{E}[W_j]|^3\right]}{\left(\sum_{j=1}^n \text{Var}(W_j)\right)^{3/2}} . \tag{138}$$

Then whenever $\sum_{j=1}^n \|x_j\|_F^2 = nTP$ and $\max_j \|x_j\|_F \le \delta n^{\frac{1}{4}}$ we have

$$B_n(x^n) \le K_1 \delta^2 \sqrt{n} + K_2 n^{1/4} + \frac{K_3}{n^{1/2}}$$

where $K_1, K_2, K_3 > 0$ are constants which only depend on channel parameters but not x^n or n.

The proof of Lemma 13 can be found in Appendix B.

D. Hypothesis testing

Many finite blocklength results are derived by considering an optimal hypothesis between appropriate distributions. We define $\beta_\alpha(P, Q)$ to be the minimum error probability of all statistical tests $P_{Z|W}$ between distributions P and Q, given that the test chooses P when P is correct with at least probability α. Formally:

$$\beta_\alpha(P, Q) = \inf_{P_{Z|W}} \left\{ \int_{\mathcal{W}} P_{Z|W}(1|w) dQ(w) : \int_{\mathcal{W}} P_{Z|W}(1|w) dP(w) \ge \alpha \right\} . \tag{139}$$

The classical Neyman-Pearson lemma shows that the optimal test achieves

$$\beta_\alpha(P, Q) = Q\left[\frac{dP}{dQ} > \gamma\right] \tag{140}$$

where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q, and γ is chosen to satisfy

$$\alpha = P\left[\frac{dP}{dQ} > \gamma\right] . \tag{141}$$

We recall a simple bound on β_α following from the data-processing inequality (see [1, (154)-(156)] or, in different notation, [27, (10.21)]):

$$\beta_\alpha(P, Q) \ge \exp\left(-\frac{D(P\|Q) + h_B(\alpha)}{\alpha}\right) . \tag{142}$$

A more precise bound [1, (102)] is

$$\beta_\alpha(P, Q) \ge \sup_{\gamma > 0} \frac{1}{\gamma}\left(\alpha - \mathbb{P}\left[\log\frac{dP}{dQ} \ge \log\gamma\right]\right) . \tag{143}$$

We will also need to define the performance of composite hypothesis tests. To this end, let $F \subset \mathcal{X}$ and $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ be a random transformation. We define

$$\kappa_\tau(F, Q_Y) = \inf_{P_{Z|Y}} \left\{ \int_\mathcal{Y} P_{Z|Y}(1|y) dQ_Y : \inf_{x \in F} \int_\mathcal{Y} P_{Z|Y}(1|y) dP_{Y|X=x} \geq \tau \right\} . \tag{144}$$

We can lower bound the error in a composite hypothesis test κ_τ by the error in an appropriately chosen binary hypothesis test as follows:

Lemma 14. *For any $P_{\tilde{X}}$ on $\mathcal{X}$ we have*

$$\kappa_\tau(F, Q_Y) \geq \beta_{\tau P_{\tilde{X}}[F]}(P_{Y|X} \circ P_{\tilde{X}}, Q_Y) \tag{145}$$

Proof. Let $P_{Z|Y}$ be any test satisfying conditions in the definition (144). We have the chain

$$\int_\mathcal{Y} P_{Z|Y}(1|y) d(P_{Y|X} \circ P_{\tilde{X}}) = \int_\mathcal{X} dP_{\tilde{X}} \int_\mathcal{Y} P_{Z|Y}(1|y) dP_{Y|X=x} \tag{146}$$

$$\geq \tau P_{\tilde{X}}[F], \tag{147}$$

where (146) is from Fubini and (147) from constraints on the test. Thus $P_{Z|Y}$ is also a test satisfying conditions in the definition of $\beta_{\tau P_{\tilde{X}}[F]}$. Optimizing over the tests completes the proof. $\square$

IV. ACHIEVABILITY

In this section, we prove the achievability side of the coding theorem for the MIMO-BF channel. We will rely on the $\kappa\beta$ bound [1, Theorem 25], quoted here:

Theorem 15 ($\kappa\beta$ bound). *Given a channel $P_{Y|X}$ with input alphabet $\mathcal{A}$ and output alphabet $\mathcal{B}$, for any distribution Q_Y on $\mathcal{B}$, any non-empty set $F \subset \mathcal{A}$, and ϵ, τ such that $0 < \tau < \epsilon < 1/2$, there exists and (M, ϵ)-max code satisfying*

$$M \geq \frac{\kappa_\tau(F, Q_Y)}{\sup_{x \in F} \beta_{1-\epsilon+\tau}(P_{Y|X=x}, Q_Y)} . \tag{148}$$

The art of applying this theorem is in choosing F and Q_Y appropriately. The intuition in choosing these is as follows: although we know the distributions in the collection $\{P_{Y|X=x}\}_{x \in F}$, we do not know which x is actually true in the composite, so if Q_Y is in the "center" of the collection, then the two hypotheses can be difficult to distinguish, making the numerator large. However, for a given x, $P_{Y|X=x}$ vs Q_Y may still be easily to distinguish, making the denominator small. The main principle for applying the $\kappa\beta$-bound is thus: Choose F and Q_Y such that $P_{Y|X=x}$ vs Q_Y is easy to distinguish for any given x, yet the composite hypothesis $Y \sim \{P_{Y|X=x}\}_{x \in F}$ is hard to distinguish from a simple one $Y \sim Q_Y$.

The main theorem of this section gives achievable rates for the MIMO-BF channel, as follows:

Theorem 16. *Fix an arbitrary caid P_X on $\mathbb{R}^{n_t \times T}$ and let*

$$V' \triangleq \frac{1}{T} \mathbb{E}\mathrm{Var}(i(X; Y, H)|X) = \mathbb{E}[V_1(X)], \tag{149}$$

where $V_1(x)$ is introduced in Proposition 8. Then we have

$$\log M^*(nT, \epsilon, P) \geq nTC(P) - \sqrt{nTV'}Q^{-1}(\epsilon) + o(\sqrt{n}) \tag{150}$$

with $C(P)$ given by (6).

Proof. Let $\tau > 0$ be a small constant (it will be taken to zero at the end). We apply the $\kappa\beta$ bound (148) with auxiliary distribution $Q_Y = (P_{Y,H}^*)^n$, where $P_{Y,H}^*$ is the caod (27), and the set F_n is to be specified

shortly. Recall notation $D_n(x^n)$, $V_n(x^n)$ and $B_n(x^n)$ from (51), (54) and (138). For any x^n such that $B_n(x^n) \leq \tau\sqrt{n}$, we have from [28, Lemma 14],

$$-\log\beta_{1-\epsilon+\tau}(P_{Y^nH^n|X^n=x^n}, P_{YH}^{*n}) \geq nTD_n(x^n) + \sqrt{nTV_n(x^n)}Q^{-1}(1-\epsilon-2\tau) - \log\frac{1}{\tau} - K' \quad (151)$$

where K' is a constant that only depends on channel parameters. We mention that obtaining (151) from [28, Lemma 14] also requires that $V_n(x^n)$ be bounded away from zero by a constant, which holds since in the expression for $V_n(x^n)$ in Proposition 8, the term (56) is strictly positive, term (57) will vanish, and terms (58) and (59) are both non-negative.

Considering (151), our choice of the set F_n should not be surprising:

$$F_n \triangleq \left\{ x^n : \|x^n\|_F^2 = nTP, V_n(x^n) \leq V' + \tau, \max_j \|x_j\|_F \leq \delta n^{\frac{1}{4}} \right\}, \quad (152)$$

where $\delta = \delta(\tau) > 0$ is chosen so that Lemma 13 implies $B_n(x^n) \leq \tau\sqrt{n}$ for any $x^n \in F_n$. Under this choice from (151), (52) and Lemma 13 we conclude

$$\sup_{x^n \in F_n} \log\beta_{1-\epsilon+\tau}(P_{Y^nH^n|X^n=x^n}, P_{YH}^{*n}) \leq -nTC(P) + \sqrt{nT(V'+\tau)}Q^{-1}(\epsilon - 2\tau) + K'', \quad (153)$$

where $K'' = K' + \log\frac{1}{\tau}$.

To lower bound the numerator $\kappa_\tau(F_n, P_{Y,H}^{*n})$ we first state two auxiliary lemmas, whose proofs follow. The first, Lemma 17, shows that the output distribution induced by an input distribution that is uniform on the sphere is "similar" (in the sense of divergence) to the n-fold product of the caod.

Lemma 17. *Fix an arbitrary caid P_X and let X^n have i.i.d. components $\sim P_X$. Let*

$$\tilde{X}^n \triangleq \frac{X^n}{\|X^n\|_F}\sqrt{nTP} \quad (154)$$

where $\|X^n\|_F = \sqrt{\sum_{t=1}^n \|X_j\|_F^2}$. Then

$$D(P_{Y^nH^n|X^n} \circ P_{\tilde{X}^n}\|P_{Y,H}^{*n}) \leq \frac{TP\log e}{n_t}\mathbb{E}\left[\|H\|_F^2\right], \quad (155)$$

where $P_{Y,H}^{*n}$ is the n-fold product of the caod (27).

The second, Lemma 18, shows that a uniform distribution on the sphere has nearly all of its mass in F_n as $n \to \infty$.

Lemma 18. *With $\tilde{X}^n$ as in Lemma 17 and set F_n defined as in (152) (with arbitrary $\tau > 0$ and $\delta > 0$) we have as $n \to \infty$,*

$$\mathbb{P}[\tilde{X}^n \in F_n] \to 1$$

Denote the right-hand side of (155) by K_1 and consider the following chain:

$$\kappa_\tau(F_n, Q_{Y^n}) \geq \exp\left(-\frac{D(P_{Y^nH^n|X^n} \circ P_{\tilde{X}^n}\|Q_{Y^n}) + \log 2}{\tau P_{\tilde{X}^n}[F_n]}\right) \quad (156)$$

$$\geq \exp\left(-\frac{K_1 + \log 2}{\tau P_{\tilde{X}^n}[F_n]}\right) \quad (157)$$

$$= \exp\left(-\frac{K_1 + \log 2}{\tau + o(1)}\right) \quad (158)$$

$$\geq K_2(\tau), \quad (159)$$

where (156) follows from Lemmas 14 and (142) with $P_{\tilde{X}^n}$ as in Lemma 17, (157) is from Lemma 17, (158) is from Lemma 18, and in (159) we introduced a τ-dependent constant K_2.

Putting (153) and (159) into the $\kappa\beta$-bound we obtain

$$\log M^*(nT, \epsilon, P) \geq nTC(P) - \sqrt{nT(V' + \tau)}Q^{-1}(\epsilon - 2\tau) - K'' - K_2(\tau).$$

Taking $n \to \infty$ and then $\tau \to 0$ completes the proof. $\qquad\square$

Now we prove the two lemmas used in the Theorem.

Proof of Lemma 17. In the case of no-fading ($H_j = 1$) and SISO, this Lemma follows from [29, Proposition 2]. Here we prove the general case. Let us introduce an auxiliary channel acting on X_j as follows:

$$\tilde{Y}_j = H_j \frac{X_j}{\|X^n\|_F}\sqrt{nTP} + Z_j, \qquad j = 1, \ldots, n \tag{160}$$

With this notation, consider the following chain:

$$D(P_{Y^n H^n | X^n} \circ P_{\tilde{X}^n} \| P^{*n}_{Y,H}) = D(P_{\tilde{Y}^n H^n | X^n} \circ P_{X^n} \| P^{*n}_{Y,H}) \tag{161}$$

$$= D(P_{\tilde{Y}^n H^n | X^n} \circ P_{X^n} \| P_{Y^n H^n | X^n} \circ P_{X^n}) \tag{162}$$

$$= D(P_{\tilde{Y}^n H^n | X^n} \| P_{Y^n H^n | X^n} | P_{X^n}) \tag{163}$$

$$= D(P_{\tilde{Y}^n | H^n, X^n} \| P_{Y^n | H^n | X^n} | P_{X^n} P_{H^n}) \tag{164}$$

$$= \frac{\log e}{2} \mathbb{E}\left[\left(1 - \frac{\sqrt{nTP}}{\|X^n\|_F}\right)^2 \sum_{t=1}^{n} \|H_j X_j\|_F^2\right] \tag{165}$$

$$= \frac{\log e}{2n_t} \mathbb{E}[\|H\|_F^2] \mathbb{E}\left[\left(\|X^n\|_F - \sqrt{nTP}\right)^2\right] \tag{166}$$

$$= \frac{\log e}{n_t} \mathbb{E}[\|H\|_F^2](nTP - \sqrt{nTP}\mathbb{E}[\|X^n\|_F]) \tag{167}$$

where (161) is by clear from (160), (162) follows since P_X is caid, (163)-(164) are standard identities for divergence, (165) follows since both $\tilde{Y}_j$ and Y_j are unit-variance Gaussians and $D(\mathcal{N}(0,1)\|\mathcal{N}(a,1)) = \frac{a^2 \log e}{2}$, (166) is from Lemma 7 and (167) is just algebra along with the assumption that $\mathbb{E}[\|X^n\|_F^2] = nTP$.

It remains to lower bound the expectation $\mathbb{E}[\|X^n\|_F]$. Notice that for any uncorrelated random variables $B_t \geq 0$ with mean 1 and variance 2 we have

$$\mathbb{E}\left[\sqrt{\frac{1}{n}\sum_{t=1}^{n} B_t}\right] \geq 1 - \frac{1}{n}, \tag{168}$$

which follows from $\sqrt{x} \geq \frac{3x - x^2}{2}$ for all $x \geq 0$ and simple computations. Next consider the chain:

$$\mathbb{E}[\|X^n\|_F] = \mathbb{E}\left[\sqrt{\sum_{i,j}\sum_{t=1}^{n}(X_t)_{i,j}^2}\right] \tag{169}$$

$$\geq \sqrt{\frac{n}{n_t T}}\sum_{i,j} \mathbb{E}\left[\sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t)_{i,j}^2}\right] \tag{170}$$

$$= \sqrt{nTP}\left(1 - \frac{1}{n}\right) \tag{171}$$

where in (171) we used the fact that for any caid, $\{(X_t)_{i,j}, t = 1, \ldots n\} \sim \mathcal{N}(0, P/n_t)$ i.i.d. (from Theorem 6) and applied (168) with $B_t = \frac{(X_t)_{i,j}^2 n_t}{P}$. Putting together (167) and (171) completes the proof. $\qquad\square$

Proof of Lemma 18. Note that since $\|X^n\|_F^2$ is a sum of i.i.d. random variables, we have $\frac{\|X^n\|_F}{\sqrt{nTP}} \to 1$ almost surely. In addition we have

$$\mathbb{E}\left[\|X_1\|_F^8\right] \leq (n_t T)^3 \sum_{i,j} \mathbb{E}\left[(X_1)_{i,j}^8\right] \triangleq K,$$

where we used the fact (Theorem 6) that X_1's entries are Gaussian. Then we have from independence of X_j's and Chebyshev's inequality,

$$\mathbb{P}[\max_j \|X_j\|_F \leq \delta' n^{\frac{1}{4}}] = \mathbb{P}[\|X_1\|_F \leq \delta' n^{\frac{1}{4}}]^n \geq \left(1 - \frac{K}{\delta'^8 n^2}\right)^n \to 1$$

as $n \to \infty$. Consequently,

$$\mathbb{P}[\max_j \|\tilde{X}_j\|_F \leq \delta n^{\frac{1}{4}}] \geq \mathbb{P}\left[\max_j \|X_j\|_F \leq \frac{\delta}{2} n^{\frac{1}{4}}\right] - \mathbb{P}\left[\frac{\|X^n\|_F}{\sqrt{nTP}} < \frac{1}{2}\right] \to 1$$

as $n \to \infty$.

Next we analyze the behavior of $V_n(\tilde{X}^n)$. From Proposition 8 we see that, due to $\|\tilde{X}^n\|_F^2 = nTP$, the term (57) vanishes, while (58) simplifies. Overall, we have

$$V_n(\tilde{X}^n) = K + \left(\frac{nTP}{\|X^n\|_F^2}\right)^2 \frac{1}{n} \sum_{j=1}^n \left(\frac{\eta_3 - \eta_4}{n_t}\|X_j\|_F^4 + \eta_4\|X_j X_j^T\|_F^2\right), \tag{172}$$

where we replaced the terms that do not depend on x^n with K. Note that the first term in parentheses (premultiplying the sum) converges almost-surely to 1, by the strong law of large numbers. Similarly, the normalized sum converges to the expectation (also by the strong law of large numbers). Overall, applying the SLLN in the limit as $n \to \infty$, we obtain:

$$\lim_{n\to\infty} V_n(\tilde{X}^n) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n V_1(\tilde{X}_j) \tag{173}$$

$$= \mathbb{E}[V_1(X)] \triangleq V'. \tag{174}$$

In particular, $\mathbb{P}[V_n(\tilde{X}^n) \leq V' + \tau] \to 1$. This concludes the proof of $\mathbb{P}[\tilde{X}^n \in F_n] \to 1$. $\qquad \square$

V. CONVERSE

Here we state and prove the converse part of Theorem 1. There are two challenges in proving the converse relative to other finite blocklength proofs. First, behavior of the information density (49) varies widely as x^n varies over the power-sphere

$$S_n = \{x^n \in (\mathbb{R}^{n_t \times T})^n : \|x^n\|_F^2 = nTP\}. \tag{175}$$

Indeed, when $\max_j \|x_j\|_F \geq cn^{\frac{1}{4}}$ the distribution of information density ceases to be Gaussian. In contrast, the information density for the AWGN channel is constant over S_n.

Second, assuming asymptotic normality, we have for any $x^n \in S_n$:

$$-\log \beta_{1-\epsilon}(P_{Y^n H^n|X^n=x^n}, P_{Y,H}^{*n}) \approx nC(P) - \sqrt{nV_n(x^n)}Q^{-1}(\epsilon) + o(\sqrt{n}).$$

However, the problem is that $V_n(x^n)$ is also non-constant. In fact there exists regions of S_n where $V_n(x^n)$ is abnormally small. Thus we need to also show that no capacity-achieving codebook can live on those abnormal sets.

The main theorem of the section is the following:

Theorem 19. *For any $\delta_n \to 0$ there exists $\delta_n' \to 0$ such that any (n, M, ϵ)-max code with $\epsilon < 1/2$ and codewords satisfying $\max_{1 \leq j \leq n} \|x_j\|_F \leq \delta_n n^{\frac{1}{4}}$ has size bounded by*

$$\log M \leq nTC(P) - \sqrt{nTV(P)}Q^{-1}(\epsilon) + \delta_n'\sqrt{n}, \tag{176}$$

where $C(P)$ and $V(P)$ are defined in (7) and (8), respectively.

Proof. As usual, without loss of generality we may assume that all codewords belong to S_n as defined in (175), see [1, Lemma 39]. The maximal probability of error code size is bounded by a meta-converse theorem [1, Theorem 31], which states that for any (n, M, ϵ) code and distribution $Q_{Y^n H^n}$ on the output space of the channel,

$$\frac{1}{M} \geq \inf_{x^n} \beta_{1-\epsilon}(P_{Y^n H^n | X = x^n}, Q_{Y^n H^n}), \tag{177}$$

where infimum is taken over all codewords. The main problem is to select $Q_{Y^n H^n}$ appropriately. We do this separately for the two subcodes defined as follows. Fix arbitrary $\delta > 0$ (it will be taken to 0 at the end) and introduce:

$$\mathcal{C}_l \triangleq \mathcal{C} \cap \{x^n : V_n(x^n) \leq n(V(P) - \delta)\} \tag{178}$$
$$\mathcal{C}_u \triangleq \mathcal{C} \cap \{x^n : V_n(x^n) > n(V(P) - \delta)\}. \tag{179}$$

To bound the cardinality of $\mathcal{C}_u$, we select $Q_{Y^n H^n} = (P_{Y,H}^*)^n$ to be the n-product of the caod (27), then apply the following estimate from [28, Lemma 14], quoted here: for any $\Delta > 0$ we have

$$\log \beta_{1-\epsilon}(P_{Y^n H^n | X = x^n}, P_{Y,H}^{*n}) \geq -nD_n(x^n) - \sqrt{nV_n(x^n)}Q^{-1}\left(1 - \epsilon - \frac{B_n(x^n) + \Delta}{\sqrt{n}}\right) - \frac{1}{2}\log\frac{n}{\Delta^2}, \tag{180}$$

where D_n, V_n and B_n are given by (52), (54) and (138), respectively. We choose $\Delta = n^{\frac{1}{4}}$ and then from Lemma 13 (which relies on the assumption that $\|x_j\|_F \leq \delta n^{\frac{1}{4}}$) we get that for some constants K_1, K_2 we have for all $x^n \in \mathcal{C}_u$:

$$B_n(x^n) + \Delta \leq K_1 \delta_n^2 \sqrt{n} + K_2 n^{\frac{1}{4}} + \frac{K_3}{n^{1/2}}.$$

From (177) and (180) we therefore obtain

$$\log|\mathcal{C}_u| \leq nTC(P) - \sqrt{nT(V(P) - \delta)}Q^{-1}(\epsilon - \delta_n'') + \frac{1}{4}\log n, \tag{181}$$

where $\delta_n'' = K_1\delta_n^2 + K_2 n^{-\frac{1}{4}} \to 0$ as $n \to \infty$.

Next we proceed to bounding $|\mathcal{C}_l|$. To that end, we first state two lemmas. Lemma 20 shows that, if in addition to the power constraint $\mathbb{E}[\|X\|_F^2] \leq TP$, we also required $\mathbb{E}[V_1(X)] \leq V(P) - \delta$, then the capacity of this variance-constrained channel is strictly less than without the latter constraint.

Lemma 20. *Consider the following constrained capacity:*

$$\tilde{C}(P, \delta) \triangleq \frac{1}{T} \sup_X \left\{ I(X; Y|H) : \mathbb{E}[\|X\|_F^2] \leq TP, \mathbb{E}[V_1(X)] \leq V(P) - \delta \right\}, \tag{182}$$

where $V(P)$ is from (8) and $V_1(x)$ is from (55). For any $\delta > 0$ there exists $\tau = \tau(P, \delta) > 0$ such that $\tilde{C}(P, \delta) < C(P) - \tau$.

Remark 8. Curiously, if we used constraint $\mathbb{E}[V_1(X)] > V(P) + \delta$ instead of $\mathbb{E}[V_1(X)] \leq V(P) - \delta$ in (182), then the resulting capacity equals $C(P)$ regardless of δ.

The following Lemma shows that, with the appropriate choice of an auxiliary distribution Q_{Y^n, H^n}, the expected size of the normalized log likelihood ratio is strictly smaller than capacity, while the variance of that same ratio is upper bounded by a constant (i.e. does not scale with n).

Lemma 21. *Define the auxiliary distribution*

$$Q_{Y|H}(y|h) = \begin{cases} P^*_{Y|H}(y|h) & \|h\|_F^2 > A \\ \tilde{P}^*_{Y|H}(y|h) & \|h\|_F^2 \leq A \end{cases} \tag{183}$$

where $A > 1$ is a constant, $P^*_{Y|H}(y|h)$ is the caod for the MIMO-BF channel, and $\tilde{P}^*_{Y|H}(y|h)$ is the caod for the variance-constrained channel in (182). Let $Q_{Y,H} = P_H Q_{Y|H}$, and $Q_{Y^n,H^n} = \prod_{i=1}^n Q_{Y,H}$. Then there exists constants $\tau, K > 0$ such that for all $x^n \in \mathcal{C}_l$,

$$C_n \triangleq \frac{1}{nT} \mathbb{E}\left[\log \frac{P_{Y^n,H^n|X^n}}{Q_{Y^n,H^n}}(Y^n, H^n|x^n)\right] \leq C(P) - \tau \tag{184}$$

$$V_n \triangleq \frac{1}{nT} \mathrm{Var}\left(\log \frac{P_{Y^n,H^n|X^n}}{Q_{Y^n,H^n}}(Y^n, H^n|x^n)\right) \leq K \tag{185}$$

where $Y_i = H_i x_i + Z_i$, $i = 1, \ldots, n$ is the joint distribution.

Remark 9. The reason we let $Q_{Y|H}$ take on two distributions depending on the value of H is because we do not know the form of $\tilde{P}^*_{Y|H}$, hence we do not explicitly know how it depends on H. This choice of $Q_{Y|H}$ ensures that expectations involving $\tilde{P}^*_{Y|H}$ are finite.

Choose $Q_{Y,H}$ as in Lemma 21, so that the bounds on C_n, V_n from (184), (185) respectively, hold. Applying [28, Lemma 15] with $\alpha = 1 - \epsilon$ (the statement of this lemma is the contents of (186)), we obtain

$$\log \beta_{1-\epsilon}(P_{Y^n,H^n|X^n=x^n}, \tilde{P}^{*n}_{Y,H}) \geq -nTC_n - \sqrt{\frac{2nTV_n}{1-\epsilon}} - \log\frac{1-\epsilon}{2} \tag{186}$$

$$\geq -nT(C(P) - \tau) - \sqrt{\frac{2nTK}{1-\epsilon}} + \log\frac{1-\epsilon}{2}. \tag{187}$$

Therefore, from (177) we conclude that for all $n \geq n_0(\delta)$ we have

$$\log |\mathcal{C}_l| \leq nT\left(C(P) - \frac{\tau}{2}\right). \tag{188}$$

Overall, from (181) and (188) we get (due to arbitrariness of δ) the statement (176). $\qquad\square$

Proof of Lemma 20. Introduce the following set of distributions:

$$\mathcal{P}' \triangleq \left\{P_X : \mathbb{E}[\|X\|_F^2] \leq TP, \ \mathbb{E}[V_1(X)] \leq V - \delta\right\}. \tag{189}$$

By Prokhorov's criterion (e.g. [30, Theorem 5.1], tightness implies relative compactness), the norm constraint implies that this set is relatively compact in the topology of weak convergence. So there must exist a sequence of distributions $\tilde{P}_n \in \mathcal{P}'$ s.t. $\tilde{P}_n \overset{w}{\to} \tilde{P}$ and $I(\tilde{X}_n; H\tilde{X}_n + Z|H) \to \tilde{C}(P, \delta)$ where $\tilde{X}_n \sim \tilde{P}_n$. By Skorokhod representation [30, Theorem 6.7], we may assume $\tilde{X}_n \overset{a.s.}{\to} \tilde{X} \sim \tilde{P}$, i.e. there exists random variable $\tilde{X}$ that is the pointwise limit of the $\tilde{X}_n$'s. Notice that for any continuous bounded function $f(h, y)$ we have

$$\mathbb{E}\left[f(H, H\tilde{X}_n + Z)\right] \to \mathbb{E}\left[f(H, H\tilde{X} + Z)\right],$$

and therefore $P_{\tilde{Y}_n,H} \overset{w}{\to} P_{\tilde{Y},H}$. Assume (to arrive at a contradiction) that $\tilde{C}(P, \delta) = C(P)$, then by the golden formula, cf. [25, Theorem 3.3], we have

$$I(\tilde{X}_n; H\tilde{X}_n + Z|H) = D(P_{YH|X}\|P^*_{Y,H}|P_{\tilde{X}_n}) - D(P_{\tilde{Y}_n,H}\|P^*_{Y,H}) \tag{190}$$

$$= \mathbb{E}\left[D_1(\tilde{X}_n)\right] - D(P_{\tilde{Y}_n,H}\|P^*_{Y,H}) \tag{191}$$

$$\leq C(P) - D(P_{\tilde{Y}_n,H}\|P^*_{Y,H}), \tag{192}$$

where $D_1(x)$ is from (52). Therefore, we have

$$D(P_{\tilde{Y}_n,H} \| P^*_{Y,H}) \to 0 \,.$$

From weak lower-semicontinuity of divergence [25, Theorem 3.6] we have $D(P_{\tilde{Y},H} \| P^*_{Y,H}) = 0$. In particular, if we denote X^* to have Telatar distribution (26), we must have

$$\mathbb{E}\left[\|\tilde{Y}\|_F^2\right] = \mathbb{E}\left[\|H\tilde{X} + Z\|_F^2\right] = \mathbb{E}\left[\|HX^* + Z\|_F^2\right]. \tag{193}$$

From Lemma 7 we have

$$\mathbb{E}\left[\|Hx\|_F^2\right] = \frac{\mathbb{E}\left[\|H\|_F^2\right]}{n_t} \|x\|_F^2 \tag{194}$$

and hence from independence of Z from (H, X) we get

$$\mathbb{E}\left[\|H\tilde{X} + Z\|_F^2\right] = \frac{\mathbb{E}\left[\|H\|_F^2\right]}{n_t} \mathbb{E}\left[\|\tilde{X}\|_F^2\right] + n_r T \,,$$

and similarly for the right-hand side of (193). We conclude that

$$\mathbb{E}\left[\|\tilde{X}\|_F^2\right] = \mathbb{E}\left[\|X^*\|_F^2\right] = TP \,.$$

Finally, plugging this fact into the expression for $D_1(x)$ in (52) and (191) we obtain

$$I(\tilde{X}; H\tilde{X} + Z | H) = \mathbb{E}\left[D_1(\tilde{X}_n)\right] = C(P) \,.$$

That is, $\tilde{X}$ is caid. But from Fatou's lemma we have (recall that $V_1(x) \geq 0$ since it is a variance)

$$\mathbb{E}\left[V_1(\tilde{X})\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[V_1(\tilde{X}_n)\right] \leq V(P) - \delta \,,$$

where the last step follows from $\tilde{P}_n \in \mathcal{P}'$. A caid achieving conditional variance strictly less than $V(P)$ contradicts the definition of $V(P)$, cf. (8), as the infimum of $\mathbb{E}\left[V_1(X)\right]$ over all caids. $\qquad\square$

Proof of Lemma 21. First we analyze C_n from (184). Denote

$$i(x; y, h) = \log \frac{P_{Y|H,X}}{P^*_{Y|H}}(y|h, x) \tag{195}$$

$$\tilde{i}(x; y, h) = \log \frac{P_{Y|H,X}}{\tilde{P}^*_{Y|H}}(y|h, x) \,. \tag{196}$$

Here, $i(x; y, h)$ is the information density given by (49), while $\tilde{i}(x; y, h)$ instead has the caod for the variance-constrained channel (182) in the denominator. Since $Q_{Y|H}$ takes on one of two distribution based on the value of H, conditioning on H in two ways yields

$$C_n = \frac{1}{nT}\mathbb{E}\left[\log \frac{P_{Y^n,H^n|X^n}}{Q_{Y^n,H^n}}(Y^n, H^n|x^n)\right] \tag{197}$$

$$= \frac{1}{nT}\sum_{j=1}^n \mathbb{E}\left[i(x_j; Y_j, H_j)\big|\|H_j\|_F^2 > A\right]\mathbb{P}[\|H_j\|_F^2 > A] \tag{198}$$

$$+ \frac{1}{nT}\sum_{j=1}^n \mathbb{E}\left[\tilde{i}(x_j; Y_j, H_j)\big|\|H_j\|_F^2 \leq A\right]\mathbb{P}[\|H_j\|_F^2 \leq A] \,. \tag{199}$$

The H_j's are i.i.d. according to P_H, so we define $p \triangleq \mathbb{P}[\|H_j\|_F^2 > A]$. Using capacity saddle point, (198) is bounded by

$$\frac{p}{nT}\mathbb{E}\left[\sum_{j=1}^n i(x_j; Y_j, H_j)\bigg|\|H_j\|_F^2 > A\right] \leq pC(P_{H>A}) \tag{200}$$

where $C(P_H)$ denotes the capacity of the MIMO-BF channel with fading distribribution P_H, and $P_{H>A}$ denotes the distribution of H conditioned on $\|H\|_F^2 > A$ (similarly, $P_{H\leq A}$ will denote H conditioned on $\|H\|_F^2 \leq A$). (200) follows from the fact that the information density, i.e. $\log \frac{P_{Y|H,X}}{P_{Y|H}^*}(y|h,x)$, is not a function of P_H, hence changing the distribution P_H does not affect the form of $i(x;y,h)$. Similarly, using Lemma 20, (199) is bounded by

$$\frac{1-p}{nT}\mathbb{E}\left[\sum_{j=1}^n \tilde{i}(X_j;Y_j,H_j)\middle|\|H_j\|_F^2 \leq A\right] \leq (1-p)\tilde{C}(P_{H\leq A}) \tag{201}$$

$$= (1-p)C(P_{H\leq A}) - \tau') \tag{202}$$

where $\tau' > 0$ is a positive constant. Putting together (200) and (202), we obtain an upper bound on C_n,

$$C_n \leq pC(P_{H>A}) + (1-p)(C(P_{H\leq A}) - \tau'). \tag{203}$$

By the law of total expectation, this simplifies to

$$C_n \leq C(P_H) - (1-p)\tau'. \tag{204}$$

Finally, we can upper bound p using Markov's inequality as

$$p = \mathbb{P}[\|H_1\|_F^2 > A] \leq \frac{1}{A} \tag{205}$$

since $A > 1$. Applying this bound to (204), we obtain

$$C_n \leq C(P_H) - (1-p)\tau' \tag{206}$$

$$\leq C(P_H) - \left(1 - \frac{1}{A}\right)\tau'. \tag{207}$$

Defining $\tau \triangleq (1 - 1/A)\tau'$ completes the proof of (184).

Next we analyze V_n from (185). The strategy will be to decompose (185) into two terms depending on the value of $\|H\|_F^2$, then show that each terms is upper bounded by $C_1 + C_2 \sum_{j=1}^n \|x_j\|_F^4$, where C_1, C_2 are constants not depending on x^n. Finally, we will show that $\sum_{j=1}^n \|x_j\|_F^4 = O(n)$ when $x^n \in \mathcal{C}_l$. To this end,

$$V_n = \frac{1}{nT}\text{Var}\left(\log \frac{P_{Y^n,H^n|X^n}}{Q_{Y^n,H^n}}(Y^n, H^n|x^n)\right) \tag{208}$$

$$= \frac{1}{nT}\sum_{j=1}^n \text{Var}\left(\log \frac{P_{Y,H|X}}{Q_{Y,H}}(Y_j, H_j|x_j)\right) \tag{209}$$

$$\leq \frac{1}{nT}\sum_{j=1}^n \mathbb{E}\left[\left(\log \frac{P_{Y,H|X}}{Q_{Y,H}}(Y_j, H_j|x_j)\right)^2\right] \tag{210}$$

where (209) follows from the independence of the terms, and (210) is from the bound $\text{Var}(X) \leq \mathbb{E}[X^2]$. Again we condition on H in two ways,

$$V_n \leq \frac{p}{nT}\sum_{j=1}^n \mathbb{E}\left[i(x_j;Y_j;H_j)^2\middle|\|H_j\|_F^2 > A\right] \tag{211}$$

$$+ \frac{1-p}{nT}\sum_{j=1}^n \mathbb{E}\left[\tilde{i}(x_j;Y_j,H_j)^2\middle|\|H_j\|_F^2 \leq A\right]. \tag{212}$$

27

For the first term, (211), we know the expression for $i(x; y, h)$ from (49), so we simply upper bound $i(x; y, h)^2$. To this end,

$$i(x; y, h)^2 \leq 2 \left(\frac{T}{2} \log \det \left(I_{n_r} + \frac{P}{n_t} h h^T \right) \right)^2 + 2 \left(\frac{\log e}{2} \sum_{j=1}^{n_{\min}} \frac{\lambda_j^2 \|v_j^T x\|^2 + 2\lambda_j \langle v_j^T x, \tilde{z}_j \rangle - \frac{P}{n_t} \lambda_j^2 \|\tilde{z}_j\|^2}{1 + \frac{P}{n_t} \lambda_j^2} \right)^2 \tag{213}$$

$$\leq C_1 \|h\|_F^2 + C_2 \|x\|_F^4 + C_3(\tilde{z}_j) \|x\|_F^2 + C_4(\tilde{z}_j) \tag{214}$$

where C_1, C_2 are non-negative constants, and $C_3(\tilde{z}_j), C_4(\tilde{z}_j)$ are functions of only $\tilde{z}_j$ that have bounded moments. This follows from:

- Bounding the first term via

  $$\left(\frac{T}{2} \log \det \left(I_{n_r} + \frac{P}{n_t} h h^T \right) \right)^2 \leq \log^2(e) \frac{P T^2}{4 n_t} n_{\min} \|h\|_F^2, \tag{215}$$

  which can be derived from the basic inequality $\log(1 + x) \leq \log(e) \sqrt{x}$.
- Noting that the second term is bounded in h, since for all $\lambda \in \mathbb{R}$,

  $$\frac{|\lambda|}{1 + \frac{P}{n_t} \lambda^2} \leq \frac{1}{2\sqrt{\frac{P}{n_t}}} \tag{216}$$

  $$\frac{\lambda^2}{1 + \frac{P}{n_t} \lambda^2} \leq \frac{n_t}{P}. \tag{217}$$

- Noting that all moments of $\|\tilde{z}_j\|^2$ are finite because this is the norm of a standard normal vector.

Therefore, after taking the expectation of (214) and summing over all n, we obtain

$$\frac{p}{nT} \sum_{j=1}^{n} \mathbb{E}\left[i(x_j; Y_j; H_j)^2 \big| \|H_j\|_F^2 > A \right] \leq \frac{1}{nT} \left(C_5 + C_6 \sum_{j=1}^{n} \|x_j\|_F^4 \right) \tag{218}$$

for some non-negative constants C_5, C_6.

To bound the second term, (212), first we split the logarithm as

$$\mathbb{E}\left[\tilde{i}(x_j; Y_j, H_j)^2 \big| \|H_j\|_F^2 \leq A \right] \tag{219}$$

$$\leq 2\mathbb{E}\left[\log \left(P_{Y|H,X}(Y_j|H_j, x_j) \right)^2 \big| \|H_j\|_F^2 \leq A \right] + 2\mathbb{E}\left[\log \left(\tilde{P}_{Y|H}^*(Y_j|H_j) \right)^2 \big| \|H_j\|_F^2 \leq A \right] \tag{220}$$

The first term in (220) is simple to handle, since its expression is given by the definition of the channel,

$$\mathbb{E}\left[\log \left(P_{Y|H,X}(Y_j|H_j, x_j) \right)^2 \big| \|H_j\|_F^2 \leq A \right] = \mathbb{E}\left[\left(-\frac{n_r T}{2} \log(2\pi) - \frac{1}{2} \|Z_j\|_F^2 \right)^2 \right] \tag{221}$$

$$\leq \frac{1}{2} n_r T \log^2(2\pi) + \frac{1}{2} n_r T (2 + n_r T) \tag{222}$$

$$\triangleq K_1 \tag{223}$$

i.e. we have a constant upper bound. For the second term in (220), notice that $\tilde{P}_{Y,H}^*$ that is inducible through channel, i.e. there exists an input distribution P_X such that $\tilde{P}_{Y,H}^*(y, h) = \mathbb{E}[P_{Y,H|X}(y, h|X)]$. Using this fact, we obtain the bound

$$-\log \tilde{P}_{Y|H}^*(y|h) = -\log \mathbb{E}[P_{Y|H,X}(y|h, X)] \tag{224}$$

$$\leq \mathbb{E}[-\log P_{Y|H,X}(y|h, X)] \tag{225}$$

$$= \mathbb{E}\left[\frac{n_r T}{2} \log(2\pi) + \frac{1}{2} \|y - hX\|_F^2 \right] \tag{226}$$

$$\leq \frac{n_r T}{2} \log(2\pi) + \|y\|_F^2 + TP \|h\|_F^2 \tag{227}$$

where (225) follows from Jensen's inequality, (226) is from the definition of the channel, and (227) follows from applying the inequality $\|A + B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$ along with $\|hX\|_F^2 \le \|h\|_F^2\|X\|_F^2$, then noting that X satisfies $\mathbb{E}[\|X\|_F^2] = TP$. Using this, we can bound the second term in (220) via

$$\mathbb{E}\left[\log\left(\tilde{P}_{Y|H}^*(Y_j|H_j)\right)^2 \Big| \|H_j\|_F^2 \le A\right] \tag{228}$$

$$\le \mathbb{E}\left[\left(\frac{n_r T}{2}\log(2\pi) + \|Y_j\|_F^2| + TP\|H_j\|_F^2\right)^2 \Big| \|H_j\|_F^2 \le A\right] \tag{229}$$

$$\le \mathbb{E}\left[3\frac{n_r^2 T^2}{4}\log^2(2\pi) + 3\|Y_j\|_F^4 + 3T^2 P^2\|H_j\|_F^4 \Big| \|H_j\|_F^2 \le A\right] \tag{230}$$

$$\le K_2 + K_3\|x\|_F^4 \tag{231}$$

where K_2, K_3 are non-negative constants which do not depend on x, (229) is from the above bound (227), and (231) follows from applying the bound

$$\mathbb{E}[\|Y_j\|_F^4|\|H_j\|_F^2 \le A] = \mathbb{E}[\|H_j x_j + Z_j\|_F^4|\|H_j\|_F^2 \le A] \tag{232}$$

$$\le 8\mathbb{E}[\|H_j\|_F^4|\|H_j\|_F^2 \le A]\|x_j\|_F^4 + 16n_r^2 T^2 \tag{233}$$

$$\le 8A\|x_j\|_F^4 + 16n_r^2 T^2. \tag{234}$$

Putting together (231) and (223), we obtain an upper bound on (212),

$$\frac{1-p}{nT}\sum_{j=1}^n \mathbb{E}\left[\tilde{i}(x_j; Y_j, H_j)^2|\|H_j\|_F^2 \le A\right] \le \frac{2(1-p)}{nT}\left(K_3 + K_4 + K_5\sum_{j=1}^n \|x_j\|_F^4\right). \tag{235}$$

Now, since $x^n \in \mathcal{C}_l$ by assumption, we can control the quantity $\sum_{i=1}^n \|x_i\|_F^4$ via

$$\sum_{i=1}^n \|x_i\|_F^4 \le \sum_{i=1}^n V_1(x_i) \tag{236}$$

$$\le n(V(P) - \delta), \tag{237}$$

where the first inequality follows from the non-negativity of the terms in $V_1(x)$ given in Proposition 8, and the second inequality is from the definition of $\mathcal{C}_l$. Hence the sum of fourth powers of the $\|x_i\|_F$'s is $O(n)$ on $\mathcal{C}_l$. All together, combining (235) and (218) yields the following bound on V_n,

$$V_n \le \frac{1}{n}\left(K' + K''\sum_{j=1}^n \|x_i\|_F^4\right) \tag{238}$$

$$\le K \tag{239}$$

which completes the proof of (185). $\qquad\square$

VI. THE RANK 1 CASE

When H is rank 1, for example in the MISO case, i.e. $n_t > n_r = 1$, the MIMO-BF channel has multiple input distributions that achieve capacity, as shown in Theorem 6. Theorem 1 proved that the dispersion in the general MIMO-BF channel is given by (8), where we minimize the conditional variance of the information density over the set of caids. In this section, we analyze those minimizers for the rank 1 case, which turns out to be non-trivial.

From Theorem 3, when H is rank 1, the conditional variance takes the form

$$V(P) = K_1 - K_2 v^*(n_t, T) \tag{240}$$

where $K_1, K_2 > 0$ are constants that depend on the channel parameters but not the input distribution. From (18), computing $v^*(n_t, T)$ requires us to maximize the variance of the squared Frobenius norm of the input distribution over the set of caids. Intuitively, this says that minimizing the dispersion is equivalent to maximizing the amount of correlation amongst the entries of X when X is jointly Gaussian. In a sense, this asks for the capacity achieving input distribution having the least amount of randomness.

Here we characterize $v^*(n_t, T)$. The manifold of caids is not easy to optimize over, since one must account for all the independence constraints on the rows and columns, the covariance constraints on the 2×2 minors, positive definite constraints, etc. as described in Theorem 6. Our strategy instead will be to give an upper bound on $v^*(n_t, T)$, then show that for certain pairs (n_t, T), the upper bound is tight. Before stating the main theorem of the section, we review orthogonal designs, which will play a large role in the solution to this problem.

A. Orthogonal designs

Definition 1 (Orthogonal Design). A real $n \times n$ orthogonal design of size k is defined to be an $n \times n$ matrix A with entries given by linear forms in $x_1, \ldots, x_k$ and coefficients in $\mathbb{R}$ satisfying

$$A^T A = \left(\sum_{i=1}^{k} x_i^2 \right) I_n \tag{241}$$

In other words, all columns of A have squared Euclidean norm $\sum_{i=1}^{k} x_i^2$, and all columns are pairwise orthogonal. A common representation for an orthogonal design is the sum $A = \sum_{i=1}^{k} x_i V_i$ where $\{V_1, \ldots, V_k\}$ is a collection of $n \times n$ real matrices satisfying Hurwitz-Radon conditions (19)-(20). Such collection is called a Hurwitz-Radon family. Theorem 4 shows that the maximal cardinality of a Hurwitz-Radon family is the *Hurwitz-Radon number* $\rho(n)$, cf. (21).

The definition of orthogonal designs can be generalized to rectangular matrices [9], as follows:

Definition 2 (Generalized Orthogonal Design). A generalized orthogonal design is a $p \times n$ matrix A with $p \geq n$ with entries as linear forms of the indeterminates $\{x_1, \ldots, x_k\}$ satisfying (241).

The quantity $R = k/p$ is often called the rate of the generalized orthogonal design. This term is justify by noticing that if p represents a number channel uses and k represents the number of data symbols, then R represents sending k data symbols in p channel uses. In this work, we are only interested in the case $R = 1$ (i.e. $k = p$), called *full-rate orthogonal designs*. Full-rate orthogonal design can be constructed from a Hurwitz-Radon family $\{V_1, \ldots, V_n\}$, each $V_i \in \mathbb{R}^{k \times k}$ by forming the matrix A

$$A = [V_1 x \ \cdots \ V_n x] \tag{242}$$

where $x = [x_1, \ldots, x_k]^T$ is the vector of indeterminates. It follows immediately from this construction that (241) is satisfied. Theorem 4 allows us to conclude that a generalized full rate $n \times k$ orthogonal design exists if and only if $n \leq \rho(k)$.

The following proposition shows that full rate orthogonal designs correspond to caids in the MIMO-BF channel.

Proposition 22. *Take $n_t = \rho(T)$ and a maximal Hurwitz-Radon family $\{V_i, i = 1, \ldots, n_t\}$ of $T \times T$ matrices (cf. Theorem 4). Let $\xi \sim \mathcal{N}(0, P/n_t I_T)$ be an i.i.d. row-vector. Then the input distribution*

$$X = \left[V_1^T \xi^T \ \cdots V_{n_t}^T \xi^T \right]^T \tag{243}$$

achieves capacity for any MIMO-BF channel provided $\mathbb{P}[\operatorname{rank} H \leq 1] = 1$.

Proof. Since $\{V_1, \ldots, V_{n_t}\}$ is a Hurwitz-Radon family, they satisfy (19)-(20). Form X as in (243). Then each row and column is jointly Gaussian, and applying the caid conditions (31) and (32) from Theorem 6 shows,

$$\mathbb{E}[R_i^T R_i] = V_i^T \mathbb{E}[\xi^T \xi] V_i = \frac{P}{n_t} V_i^T V_i = \frac{P}{n_t} I_T \tag{244}$$

$$\mathbb{E}[R_i^T R_j] = V_i^T \mathbb{E}[\xi^T \xi] V_j = \frac{P}{n_t} V_i^T V_j = -\frac{P}{n_t} V_j^T V_i = -\mathbb{E}[R_j^T R_i] \tag{245}$$

Therefore X satisfies the caid conditions, and hence achieves capacity. $\qquad \square$

Remark 10. The above argument implies that if $X \in \mathbb{R}^{n_t \times T}$ is constructed above, then removing the last row of X gives an $(n_t - 1) \times T$ input distribution that also achieves capacity.

B. Proof of theorem 5

Theorem 5 states that for dimensions where orthogonal designs exist, the conditional variance (8) is minimized if and only if the input is constructed from an orthogonal design as in Proposition 22. The approach is first to prove an upper bound on v^*, then show that conditions for tightness of the upper bound correspond to conditions of the Hurwitz-Radon theorem.

We start with a simple lemma, which will be applied with A, B equal to the rows of the capacity achieving input X.

Lemma 23. *Let $A = (A_1, \ldots, A_n)$ and $B = (B_1, \ldots, B_n)$ each be i.i.d. random vectors from the same distribution with finite second moment $\mathbb{E}[A_1^2] = \sigma^2 < \infty$. While A and B are i.i.d. individually, they may have arbitrary correlation between them. Then*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(A_i, B_j) \leq n\sigma^2 \tag{246}$$

with equality iff $\sum_{i=1}^{n} A_i = \sum_{i=1}^{n} B_i$ almost surely.

Proof. Simply use the fact that covariance is a bilinear function, and apply the Cauchy-Schwarz inequality as follows:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(A_i, B_j) = \mathrm{Cov}\left(\sum_{i=1}^{n} A_i, \sum_{j=1}^{n} B_j\right) \tag{247}$$

$$\leq \sqrt{\mathrm{Var}\left(\sum_{i=1}^{n} A_i\right) \mathrm{Var}\left(\sum_{j=1}^{n} B_j\right)} \tag{248}$$

$$= \sqrt{(n\mathrm{Var}(A_1))(n\mathrm{Var}(B_1))} \tag{249}$$

$$= n\sigma^2 \tag{250}$$

We have equality in Cauchy-Schwarz when $\sum_{i=1}^{n} A_i$ and $\sum_{i=1}^{n} B_i$ are proportional, and since these sums have the same distribution, the constant of proportionality must be equal to 1, so we have equality in (246) iff $\sum_{i=1}^{n} A_i = \sum_{i=1}^{n} B_i$ almost surely. $\qquad \square$

Proof of Theorem 5. First, we rewrite $v^*(n_t, T)$ defined in (18) as

$$v^*(n_t, T) \triangleq \frac{n_t^2}{2P^2} \max_{P_X : I(X;Y|H)=C} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \sum_{k=1}^{T} \sum_{l=1}^{T} \mathrm{Cov}(X_{i,k}^2, X_{j,l}^2) \tag{251}$$

From here, $v^*(n_t, T) = v^*(T, n_t)$ follows from the symmetry to transposition of the caid-conditions on X (see Theorem 6) and symmetry to transposition of (251). From now on, without loss of generality we assume $n_t \leq T$.

For the upper bound, since the rows and columns of X are i.i.d., we can apply Lemma 23 with $A_i = X_{i,k}^2$ and $B_j = X_{j,l}^2$ (and hence $\sigma^2 = 2(P/n_t)^2$) to get

$$\sum_{i,j,k,l} \text{Cov}(X_{i,k}^2, X_{j,l}^2) \leq \sum_{i,j} 2T(P/n_t)^2 = 2n_t^2 T(P/n_t)^2 \, , \tag{252}$$

which together with (251) yields the upper bound (22) (recall that $n_t \leq T$).

Equation (252) implies that if X achieves the bound (22), then removing the last row of X achieves (22) as an $(n_t - 1) \times T$ design. In other words, if (22) is tight for $n_t \times T$ then it is tight for all $n_t' \leq n_t$.

Notice that for any X such that any pair $X_{i,k}, X_{j,l}$ is jointly Gaussian, we have

$$\frac{n_t^2}{2P^2} \text{Var}(\|X\|_F^2) = \sum_{i,j,k,l} \rho_{ikjl}^2 \, , \tag{253}$$

where

$$\rho_{ikjl} \triangleq \frac{n_t}{P} \text{Cov}(X_{ik}, X_{jl}) \, . \tag{254}$$

Take $X \in \mathbb{R}^{n_t \times T}$ as constructed in (243). By Proposition 22, X is capacity achieving and identity (253) clearly holds. In the representation (243), the matrix $V_j^T V_i$ contains the correlation coefficients between rows i and j of X, since $\mathbb{E}[(\xi V_j)^T(\xi V_i)] = \frac{P}{n_t} V_j^T V_i$, so

$$\|V_j^T V_i\|_F^2 = \sum_{k=1}^{T} \sum_{l=1}^{T} \rho_{ikjl}^2 \, . \tag{255}$$

Therefore we can represent the sum of squared correlation coefficients as

$$\sum_{i,j,k,l} \rho_{ijkl}^2 = \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \|V_j^T V_i\|_F^2 \tag{256}$$

$$= \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \text{tr}\left(V_j V_j^T V_i V_i^T\right) \tag{257}$$

$$= \text{tr}\left(\left(\sum_{i=1}^{n_t} V_i V_i^T\right)^2\right) \tag{258}$$

$$= n_t^2 . T \tag{259}$$

Line (259) follows since the V_i's are orthogonal by the Hurwitz-Radon condition, so each $V_i V_i^T = I_T$ in the summation in (258). Hence the X constructed in (243) achieves the upper bound in (252) and (22).

Next we prove (24). Suppose X is a jointly-Gaussian caid saturating the bound (252). From Lemma 23, the condition for equality in (246) implies that for all $j \in \{1, \ldots, n_t\}$,

$$\|R_j\|_F^2 = \|R_1\|_F^2 \quad a.s. \tag{260}$$

where R_j is the j-th row of X for $j = 1, \ldots, n_t$. In particular, this means that every R_j is a linear function of R_1. Consequently, we may represent X in terms of a row-vector $\xi \sim \mathcal{N}(0, P/n_t I)$ as in (243), that is $R_j = \xi V_j$ for some $T \times T$ matrices $V_j, j \in [n_t]$. We clearly have

$$\mathbb{E}[R_i^T R_j] = \frac{P}{n_t} V_i^T V_j \, .$$

But then the caid constraints (31)-(32) imply that the matrix A in (242) constructed using indeterminates $\{x_1, \ldots, x_{n_t}\}$ and family $\{V_1, \ldots, V_{n_T}\}$ satisfies Definition 2. Therefore, from Theorem 4, (see also [31, Proposition 4]), we must have $n_T \leq \rho(T)$. $\qquad \square$

Remark 11. In the case $n_t = T = 2$ it is easy to show that for any non-jointly-Gaussian caid, there exists a jointly-Gaussian caid achieving the same $\mathrm{Var}(\|X\|_F^2)$. Indeed, consider (37) with $\rho = \frac{\mathrm{cov}(X_{1,1}^2, X_{2,2}^2) + \mathrm{cov}(X_{1,2}^2, X_{2,1}^2)}{8(P/n_t)^2}$. If this phenomena held in general, we would conclude that (23) holds if and only if $n_t \leq \rho(T)$ or $T \leq \rho(n_t)$. As a step towards the proof of the latter, we notice that any caid X achieving equality in (252) satisfies

$$XX^T = \frac{\|X\|_F^2}{n_t} I_{n_t} \qquad \text{(a.s.)} , \tag{261}$$

which is equivalent to saying $R_i R_j' = 0$ for $i \neq j$. The latter follows from applying (260) to rows of UX, where U is an arbitrary orthogonal matrix. Identity (261) could be informally stated as "any caid saturating (252) is a random full-rate orthogonal design".

In summary, the full-rate orthogonal designs (when those exist) achieve the optimal channel dispersion $V(P)$. Some examples (ξ_j are i.i.d. $\mathcal{N}(0,1)$) for $n_t = T = 4$ and $n_t = 4, T = 3$, respectively, are as follows:

$$X = \sqrt{\frac{P}{4}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 & \xi_4 \\ -\xi_2 & \xi_1 & -\xi_4 & \xi_3 \\ -\xi_3 & \xi_4 & \xi_1 & -\xi_2 \\ -\xi_4 & -\xi_3 & \xi_2 & \xi_1 \end{bmatrix} \tag{262}$$

$$X = \sqrt{\frac{P}{4}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ -\xi_2 & \xi_1 & -\xi_4 \\ -\xi_3 & \xi_4 & \xi_1 \\ -\xi_4 & -\xi_3 & \xi_2 \end{bmatrix}$$

C. Beyond full-rate orthogonal designs

For pairs (n_t, T) where $n_t > \rho(T)$, full-rate orthogonal design do not exist. For example $\rho(3) = 1$, so no full-rate orthogonal design exits for $n_t = 2$, $T = 3$. Which caids are minimizer for (8) in this case? In general, we do not know the answer and do not even know whether one can restrict the search to jointly-Gaussian caids. But one thing is certain: it is definitely not an i.i.d. Gaussian (Telatar) caid. To show this claim, we will give a method for constructing improved caids.

To that end, suppose that X consists of entries $\pm \xi_j$, $j = 1 \ldots, d$, where $\xi_j \overset{i.i.d.}{\sim} \mathcal{N}(0, P/n_t)$. Then we have:

$$\frac{n_t^2}{2P} \mathrm{Var}(\|X\|_F^2) = \sum_{t=1}^{d} (\ell_t)^2 , \tag{263}$$

where ℓ_t is the number of times $\pm \xi_t$ appears in the description of X. By this observation and the remark after Theorem 6 (any submatrix of a caid X is also a caid), we can obtain lower bounds on $v^*(n_t, T)$ for $n_t > \rho(T)$ via the following *truncation construction*:

1) Take $T' > T$ such that $\rho(T') \geq n_t$ and let X' be a corresponding $\rho(T') \times T'$ full-rate orthogonal design with entries $\pm \xi_1, \ldots \pm \xi_{T'}$.
2) Choose an $n_t \times T$ submatrix of X' maximizing the sum of squares of the number of occurrences of each of ξ_j, cf. (263).

TABLE I
VALUES FOR $v^*(n_t, T)$

$n_t \setminus T$	1	2	3	4	5	6	7	8
1	1	2	3	4	5	6	7	8
2		8	10*	16	18	24	26	32
3			21*	36	[39,45]	[46,54]	[57,63]	72
4				64	[68,80]	[80,96]	[100,112]	128
5					[89,125]	[118,150]	[155,175]	200
6						[168,216]	[222,252]	288
7							[301,343]	392
8								512

Note: Table is symmetric about diagonal; intervals $[a, b]$ mark entries for which dispersion-optimal input is unknown. The optimality of entries marked with $*$ is only established in the class of all jointly-Gaussian caids.

As an example of this method, by truncating a 4×4 design (262) we obtain the following 2×3 and 3×3 submatrices:

$$X = \sqrt{\frac{P}{3}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ -\xi_2 & \xi_1 & \xi_4 \\ -\xi_3 & -\xi_4 & \xi_1 \end{bmatrix} \qquad X = \sqrt{\frac{P}{2}} \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 \\ -\xi_2 & \xi_1 & \xi_4 \end{bmatrix} \tag{264}$$

By independent methods we were able to show that designs (264) are dispersion-optimal out of all jointly Gaussian caids. Note that in these cases (23) does not hold, illustrating (24).

Our current knowledge about v^* is summarized in Table I. The lower bounds for cases not handled by Theorem 5 were computed by truncating the 8x8 orthogonal design [9, (5)]. Based on the evidence from $2 \times T$ and 3×3 we *conjecture* this construction to be optimal.

From the proof of Theorem 5 it is clear that Telatar's i.i.d. Gaussian is never dispersion optimal, unless $n_t = 1$ or $T = 1$. Indeed, for Telatar's input $\rho_{ikjl} = 0$ unless $(i, k) = (j, l)$. Thus embedding even a single 2×2 Alamouti block into an otherwise i.i.d. $n_t \times T$ matrix X strictly improves the sum (251). We note that the value of $\frac{V}{C^2}$ entering (2) can be quite sensitive to the suboptimal choice of the design. For example, for $n_t = T = 8$ and $SNR = 20\ dB$ estimate (2) shows that one needs

- around 600 channel inputs (that is 600/8 blocks) for the optimal 8×8 orthogonal design, or
- around 850 channel inputs for Telatar's i.i.d. Gaussian design

in order to achieve 90% of capacity. This translates into a 40% longer delay or battery spent in running the decoder.

Thus, curiously even in cases where pure multiplexing (that is maximizing transmission rate) is needed – as is often the case in modern cellular networks – transmit diversity enters the picture by enhancing the finite blocklength fundamental limits. Remember, however, that our discussion pertains only to cases when the transmitter (base-station) is equipped with more antennas than the receiver (user equipment), or when the channel does not have more than one diversity branch.

In cases when full-rate designs do not exist, there have been various suggestions as to what could be the best solution, e.g. [31]. Thus for non full-rate designs the property of minimizing dispersion (such as (264)) could be used for selecting the best design for cases $n_t > \rho(T)$.

VII. DISCUSSION

Figure 1 plots the capacity, normal approximation, and $\beta\beta$ achievability bound for the MIMO channel with $n_t = n_r = T = 4$ for the complex case. The details of this computation are given in [19]. The $\beta\beta$ bound was developed by Yang et al [19] and is often more computationally friendly than the $\kappa\beta$ bound. This figure illustrates the gap between achievability and the normal approximation, as well as the gap to capacity. For example, at blocklength 400, we can achieve about 88% of capacity, and at blocklength
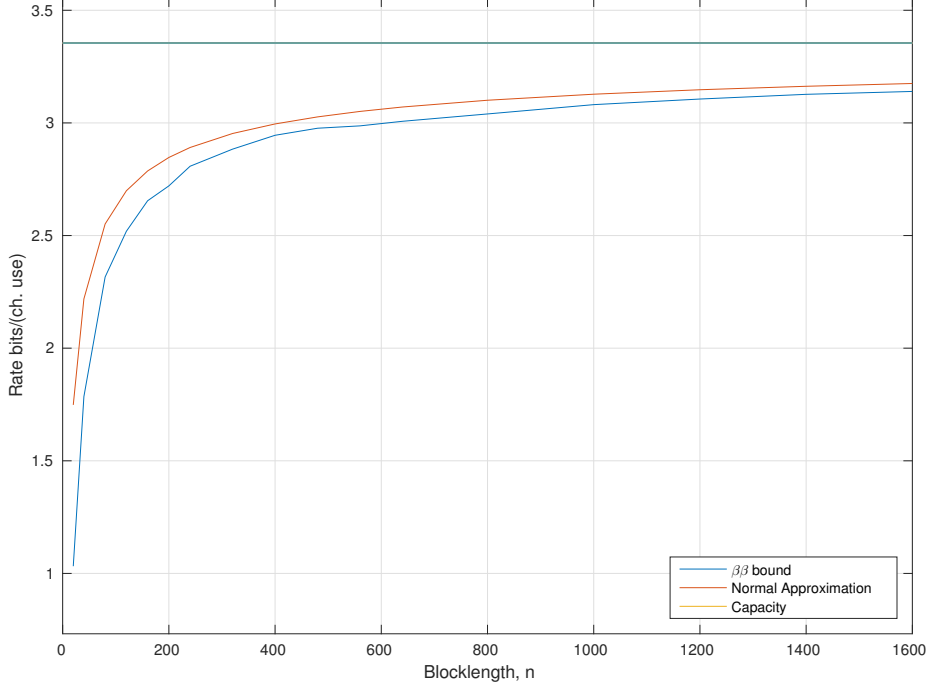
Fig. 1. Achievability and normal approximation for $n_t = n_r = T = 4$, $P = 0$dB, and $\epsilon = 10^{-3}$.

1000 we can achieve about about 92% of capacity, given $P = 0$dB and tolerating an error probability of 10^{-3}.

Figure 2 shows the dependence of the rate on the coherence time T for the 4×4 MIMO channel. The normal approximation for $T = 1, 20, 80$ is plotted. From (6) and (12), we know the capacity does not depend on T, but the dispersion depends on T in an affine relationship. Hence, from the dispersion we see that a larger coherence time reduces the maximum transmission rate when the other channel parameters are held fixed. Intuitively, when the coherence time is lower, we are able to average over independent realizations of the fading coefficients in less channel uses. Note that the CSIR assumption implies that we know the channel coefficient perfectly, which may be unrealistic at short coherence times for a practical channel.

We now ask: how does the dispersion depend on the number of transmit and receive antennas? Figures 3 and 4 depict the normalized dispersion V/C^2, cf. (2), as a function of the number of antennas. The fading process is chosen to be i.i.d. $\mathcal{N}(0, 1)$. Each plot has two curves: one curve with n_r fixed and n_t growing, and the other curve with n_t fixed and n_r growing. In both plots, coherence time is $T = 16$. The difference is that on Fig. 3 the *received power* P_r is held fixed (at 20 dB, i.e. P is chosen so that $P_r = 100$), whereas on Fig 4 it is the transmit power P that is held fixed (also at $20 \ dB$, i.e. $P = 100$). The relation between P_r and P is as follows:

$$P_r = \frac{P}{n_t}\mathbb{E}[\|H\|_F^2] \,, \tag{265}$$

These figures also display the asymptotic limiting values of $\frac{V}{C^2}$ computed via random-matrix theory:
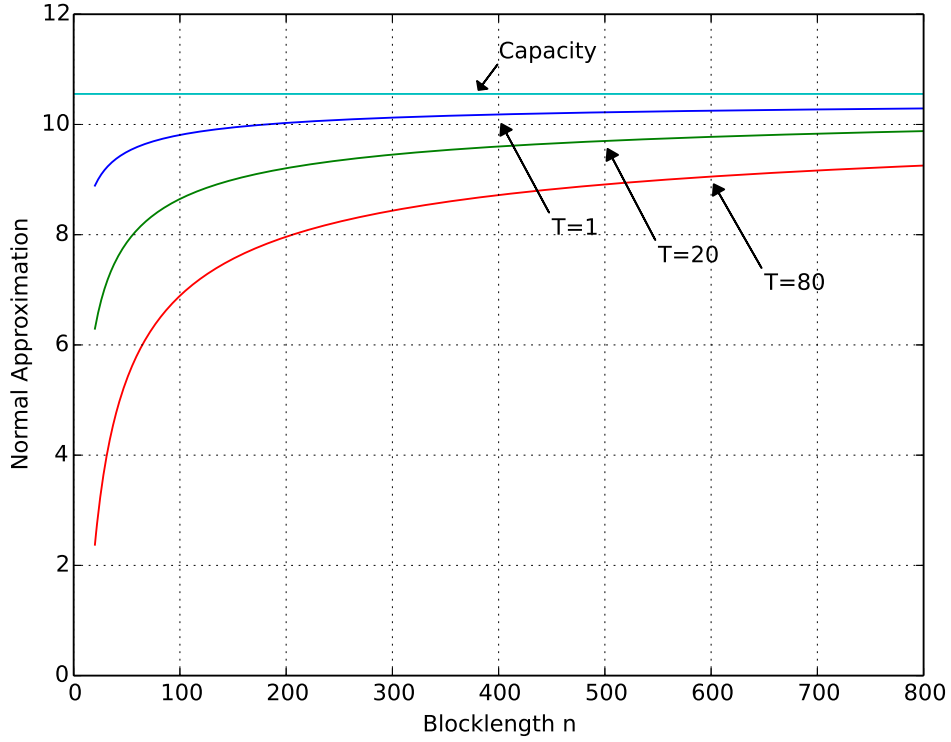
35

Fig. 2. The normal approximation for varying coherent times, with $n_t = n_r = 4$, $P = 20dB$, and $\epsilon = 10^{-3}$

1) When n_r is fixed and $n_t \to \infty$ under fixed *received power* P_r we have

$$C(P_r) = \frac{n_r}{2} \log \left(1 + \frac{P_r}{n_r} \right) + o(1) \tag{266}$$

$$V(P_r) = \log^2(e) \frac{P_r}{1 + \frac{P_r}{n_r}} + o(1) \ . \tag{267}$$

2) When n_t is fixed and $n_r \to \infty$ under fixed *received power* P_r we have

$$C(P_r) = \frac{n_t}{2} \log \left(1 + \frac{P_r}{n_t} \right) + o(1) \tag{268}$$

$$V(P_r) = \log^2(e) \frac{P_r(2 + \frac{P_r}{n_t})}{2(1 + \frac{P_r}{n_t})^2} + o(1) \ . \tag{269}$$

3) When n_r is fixed and $n_t \to \infty$ under fixed *transmitted power* P we have

$$C(P) = \frac{n_r}{2} \log \left(1 + P \right) + o(1) \tag{270}$$

$$V(P) = \log^2(e) \frac{n_r P}{1 + P} + o(1) \ . \tag{271}$$

4) When n_t is fixed and $n_r \to \infty$ under fixed *transmit power* P we have

$$C(P) = \frac{n_t}{2} \log \left(1 + \frac{n_r P}{n_t} \right) + o(1) \tag{272}$$

$$V(P) = \log^2(e) \frac{n_t}{2} + o(1) \ . \tag{273}$$
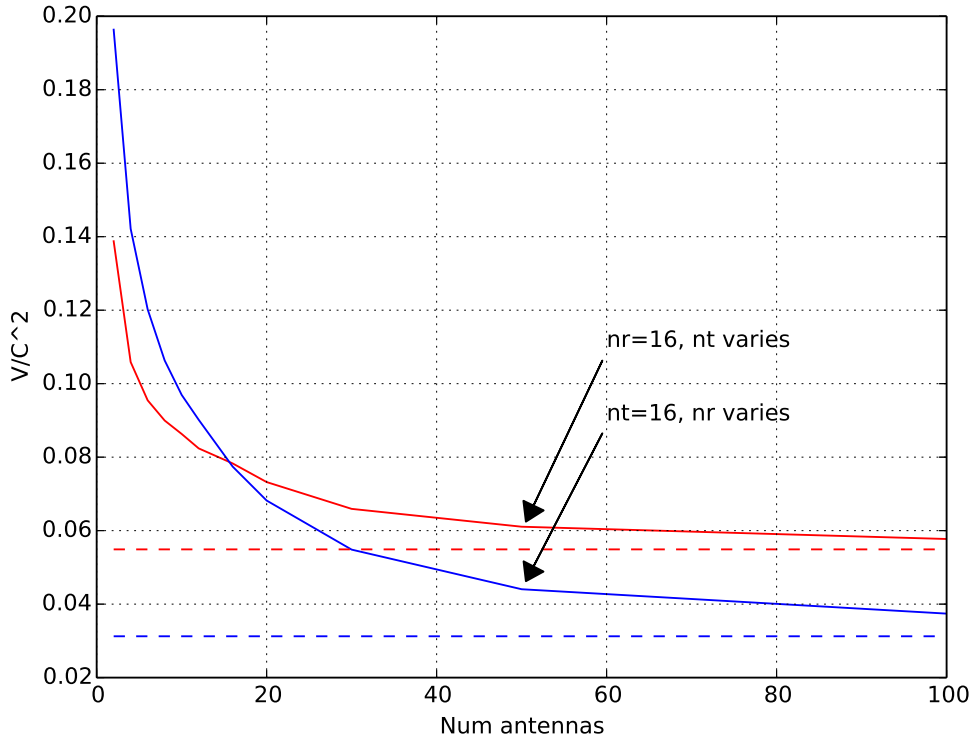
36

Fig. 3. Normalized dispersion $\frac{V}{C^2}$ as a function of n_r and n_t. The <u>received</u> power is $P_r = 20dB$ and $T = 16$. Dashed lines are asymptotic values from (266)-(269).

Note that when the received power is fixed, *reciprocity* holds: the capacity of the $n_t \times n_r$ channel is the same as the capacity of the $n_r \times n_t$ one. Having information about dispersion, we may ask the more refined question: although capacities of the channels are the same, which one has better dispersion (i.e. causes smaller coding latency)?

From approximations (267) and (269), we can see that the channel dispersion is not symmetric in n_t, n_r. For example, in the setting of Fig. 3 we see that the delay penalty in the $n_t \ll n_r$ regime is 58% of the penalty in the $n_r \ll n_t$ regime. Hence, in a two user channel, if user 1 has n_1 antennas and user 2 has $n_2 > n_1$ antennas, then the asymptotic analysis suggest that channel from user 1 to user 2 can support higher rates than the channel from user 2 to user 1 at finite blocklength.

Figure 4 shows the scenario where the transmit power is fixed. In this case, the capacity approaches a finite limit when n_r is held fixed and $n_t \to \infty$, but grows logarithmically when n_t is fixed and $n_r \to \infty$, as shown in equations (270) and (272). In this setting, the normalized dispersion approaches a finite limit when n_r is fixed and $n_t \to \infty$, yet it vanishes when n_t is fixed and $n_r \to \infty$. Consequently in this regime, we can always choose the number of receive antennas n_r large enough so that our system can achieve a given fraction of capacity η using blocklength n. The normalized dispersion in this case is proportional to $1/\log^2(n_r)$.

APPENDIX A
EXISTENCE OF NON-GAUSSIAN CAIDS

Proposition 24. *Let $S \subset \mathbb{R}^n$ be such that a) $0 \in S$ and b) there exists a non-zero polynomial in n variables with real coefficients vanishing on S. Then there exists a random variable X taking values in*
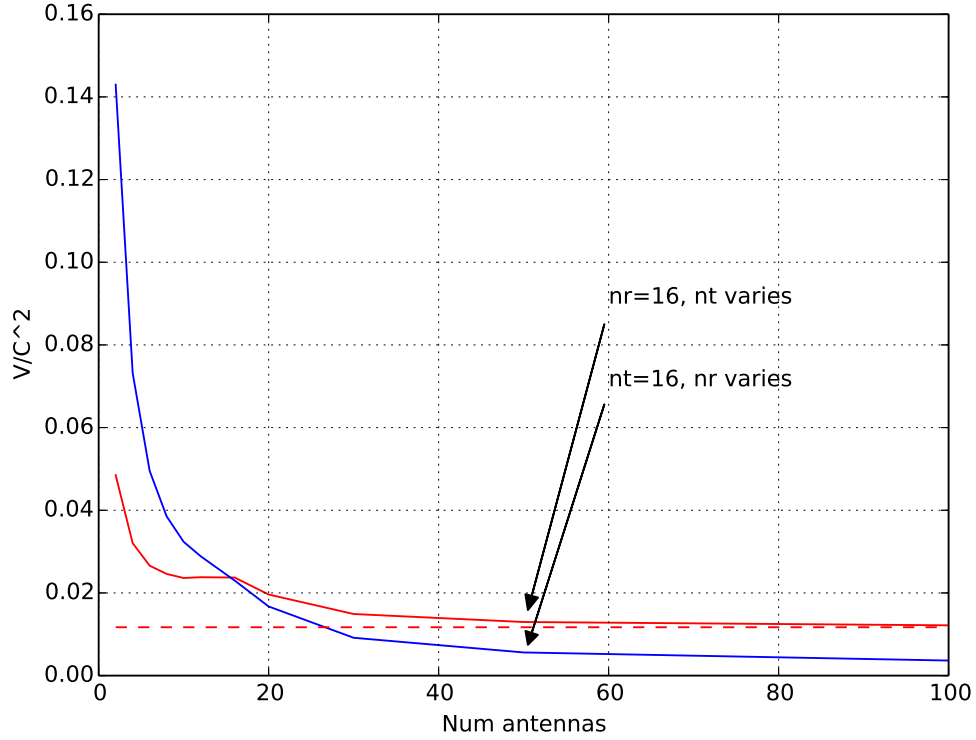
37

Fig. 4. Normalized dispersion $\frac{V}{C^2}$ as a function of n_r and n_t. The <u>transmit</u> power is $P = 20dB$ and $T = 16$. Dashed lines are asymptotic values from (270)-(273).

$\mathbb{R}^n$ *with the property that its characteristic function* $\Psi(t) \overset{\triangle}{=} \mathbb{E}\left[e^{i\sum_{k=1}^n t_j X_j}\right], t \in \mathbb{R}^n$ *satisfies*

$$\Psi(t) = e^{-\frac{\|t\|_2^2}{2}} \qquad \forall t \in S$$

but there exist a $t_0 \in \mathbb{R}^n$ *such that* $\Psi(t_0) \neq e^{-\frac{\|t_0\|_2^2}{2}}$ *(i.e.* $X \not\sim \mathcal{N}(0, I_n)$*).*

Remark 12. The simplest application of this proposition is the following. Suppose that three random vectors in $\mathbb{R}^3$ have the property that projection onto any (2-dimensional) plane has the joint distribution $\mathcal{N}(0, I_2) \times \mathcal{N}(0, I_2) \times \mathcal{N}(0, I_2)$. Does it imply that the joint distribution of them is $\mathcal{N}(0, I_3) \times \mathcal{N}(0, I_3) \times \mathcal{N}(0, I_3)$? Note that it is easy to argue that joint distribution of any pair of them is indeed $\mathcal{N}(0, I_3) \times \mathcal{N}(0, I_3)$ and thus the only *jointly Gaussian* distribution that satisfies the requirements is indeed the i.i.d. triplet. However, the above proposition shows that the general answer is still negative. Here S is a subset of all $\mathbb{R}^{3 \times 3}$ with determinant zero.

Proof. We will slightly extend the argument of [32]. We will assume familiarity with basic commutatitive algebra on the level of [33]. Consider an identity expressing the well-known computation of the Gaussian characteristic function:

$$\frac{1}{\sqrt{2\pi\alpha^2}} \int_{\mathbb{R}} e^{itx - \frac{x^2}{2\alpha^2}} = e^{-\alpha^2 \frac{t^2}{2}} .$$

Setting $\beta = \frac{1}{\alpha^2}$, changing sign of t we get

$$\int_{\mathbb{R}} e^{-itx - \frac{\beta x^2}{2}} dx = \sqrt{\frac{2\pi}{\beta}} e^{-\frac{t^2}{2\beta}} .$$

38

Differentiating this in β and setting $\beta = \frac{1}{2}$ we get

$$\int_{\mathbb{R}} x^{2k} e^{-itx - \frac{x^2}{4}} \, dx = p_{2k}(t) e^{-t^2} \,,$$

where $p_{2k}(t)$ is some polynomial of degree $2k$ with real coefficients (and involving only even powers of t). For later convenience, we also interchange t and x to get

$$\int_{\mathbb{R}} t^{2k} e^{-itx - \frac{t^2}{4}} \, dt = p_{2k}(x) e^{-x^2} \,. \tag{274}$$

(Identity (274) also follows from the fact that Hermite polynomials times Gaussian density are eigenfunctions of the Fourier transform.)

Next, suppose that there is a polynomial $q(t_1, \ldots, t_n)$ such that q vanishes on S and each monomial $t_1^{k_1} \cdots t_n^{k_n}$ in q has all $k_1, \ldots, k_n$ even. Then, define the characteristic function

$$\Psi(t_1, \ldots, t_n) \overset{\triangle}{=} e^{-\frac{\sum_{k=1}^n t_k^2}{2}} + \epsilon e^{-\frac{\sum_{k=1}^n t_k^2}{4}} q(t_1, \ldots, t_n) \,. \tag{275}$$

We will argue that for ϵ sufficiently small, Ψ is a characteristic function of some (obviously non-Gaussian) probability density function f on $\mathbb{R}^n$. By taking the inverse Fourier transform we get that

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{\|x\|_2^2}{2}} (1 + \epsilon g(x)) \,.$$

where $e^{-\frac{\|x\|_2^2}{2}} g(x)$ is the inverse Fourier transform of the second term in (275). Since $\Psi(t)$ is even in each t_j, we conclude that $f(x)$ is real. Since $q(0) = 0$ (recall that $0 \in S$) we have $\Psi(0) = 1$, and thus $\int_{\mathbb{R}^n} f = 1$. So to prove that f is a valid density function for small ϵ we only need to show that

$$\sup_{x \in \mathbb{R}^n} |g(x)| < \infty \,. \tag{276}$$

To that end, notice that applying (274) to each monomial $\prod t_j^{2k_j}$ we get

$$\int_{\mathbb{R}^n} \left(\prod_{j=1}^n t_j^{2k_j} \right) e^{-i \sum_j t_j x_j - \frac{\|t\|_2^2}{4}} \, dt_1 \cdots dt_n = \left(\prod_j p_{2k_j}(x_j) \right) e^{-\|x\|_2^2} \,. \tag{277}$$

Multiplying the right-hand side by $e^{\frac{\|x\|_2^2}{2}}$ we conclude that contribution of each monomial of q to $\sup_x |g(x)|$ is bounded by

$$\sup_{x \in \mathbb{R}^n} \left| \left(\prod_j p_{2k_j}(x_j) \right) e^{-\frac{\|x\|_2^2}{2}} \right| < \infty \,.$$

Since there are finitely many monomials in q, the proof of (276) and of validity of $\Psi(t)$ is done.

We are left to argue that there must necessarily exist polynomial q with required properties. By assumption there exist some other polynomial q_0 vanishing on S. Consider an inclusion of rings

$$T \overset{\triangle}{=} \mathbb{R}[x_1^2, x_2^2, \ldots, x_n^2] \hookrightarrow \mathbb{R}[x_1, \ldots, x_n]$$

where $\mathbb{R}[x_1, \ldots, x_n]$ denotes the ring of polynomials with variables $x_1, \ldots, x_n$ and coefficients in $\mathbb{R}$, and $\hookrightarrow$ denotes an inclusion map. This morphism of rings is obviously finite. Consider ideal (q_0) of $\mathbb{R}[x_1, \ldots, x_n]$ and denote as usual by $(q_0)^c \overset{\triangle}{=} (q_0) \cap T$ its contraction. We argue that $(q_0)^c \neq (0)$. Assume otherwise, then we have $(q_0)^c = (0)$ and $\sqrt{(q_0)}^c = (0)$ (since $\sqrt{(0)} = (0)$ as T is an integral domain). Take all minimal primes of (q_0), call these $\{\mathfrak{p}_j\}$, then the radical of (q_0) is the intersection of all prime ideals that contain it, i.e. $\sqrt{(q_0)} = \cap_j \mathfrak{p}_j$. Then, denoting $\mathfrak{q}_j \overset{\triangle}{=} \mathfrak{p}_j^c$ we get that $\cap_j \mathfrak{q}_j = (0)$ in T. By "prime-avoidance", cf. [33, Prop. 1.11], we know $(0) \subset \cap_j \mathfrak{q}_j$ implies that $\mathfrak{q}_j \subset (0)$ for some j, hence $\mathfrak{q}_j$ is the zero ideal for some j. This contradicts "going-up theorem", cf. [33, Corollary 5.9], so we must have $(q_0)^c \neq (0)$, and hence we may take q as an arbitrary non-zero element of $(q_0)^c$. $\qquad \square$

Proof of Lemma 13. We begin with upper bounding the numerator in (138), i.e.

$$\sum_{j=1}^{n} \mathbb{E}\left[|W_j - \mathbb{E}[W_j]|^3\right] . \tag{278}$$

The information density is given by

$$i(x; y, h) = \frac{1}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{j=1}^{T} \|y_j - h x_j\|^2 + \frac{1}{2} \text{tr}\left(y^T \Sigma^{-1} y\right) \tag{279}$$

where

$$\Sigma = I_{n_r} + \frac{P}{n_t} H H^T . \tag{280}$$

Define $W = i(x; Y, H)$ under the distribution $Y = Hx + Z$. (279) reduces to

$$W = \frac{T}{2} \log \det(\Sigma) - \frac{1}{2}\|Z\|_F^2 + \frac{1}{2}\text{tr}\left(x^T H^T \Sigma^{-1} H x + 2 x^T H^T \Sigma^{-1} Z + Z^T \Sigma^{-1} Z\right) \tag{281}$$

$$= c(H, Z) + \frac{1}{2}\text{tr}\left(x^T H^T \Sigma^{-1} H x\right) + \text{tr}\left(x^T H^T \Sigma^{-1} Z\right) \tag{282}$$

where the scalar random variable

$$c(H, Z) = \frac{T}{2} \log \det(\Sigma) - \frac{1}{2}\|Z\|_F^2 + \frac{1}{2}\text{tr}(Z^T \Sigma^{-1} Z) \tag{283}$$

is the sum of all the terms that do not depend on x. Note that

$$\mathbb{E}\text{tr}\left(x^T H^T \Sigma^{-1} H x\right) = \text{tr}(x^T \mathbb{E}[H^T \Sigma^{-1} H] x) \tag{284}$$

$$\mathbb{E}\text{tr}\left(x^T H^T \Sigma^{-1} Z\right) = 0 . \tag{285}$$

Therefore, the "centered" information density is

$$W - \mathbb{E}[W] = c_0(H, Z) - \mathbb{E}[c(H, Z)] + \frac{1}{2}\text{tr}\left(x^T \left(H^T \Sigma^{-1} H - \mathbb{E}[H^T \Sigma^{-1} H]\right) x\right) + \text{tr}\left(x^T H^T \Sigma^{-1} Z\right) \tag{286}$$

$$= c_0(H, Z) + \text{tr}(x^T A x) + \text{tr}(x^T B) \tag{287}$$

where

$$A = \frac{1}{2}(H^T \Sigma^{-1} H - \mathbb{E}[H^T \Sigma^{-1} H]) \tag{288}$$

$$B = H^T \Sigma^{-1} Z \tag{289}$$

$$c_0(H, Z) = c(H, Z) - \mathbb{E}[c(H, Z)] . \tag{290}$$

Hence we can upper bound the centered third moment as

$$\mathbb{E}[|W - \mathbb{E}[W]|^3] \leq 3 \underbrace{\mathbb{E}[|c_0(H, Z)|^3}_{S_1} + 3 \underbrace{\mathbb{E}[|\text{tr}(x^T A x)|^3]}_{S_2} + 3 \underbrace{\mathbb{E}[|\text{tr}(x^T B)|^3]}_{S_3} . \tag{291}$$

We now proceed to upper bound each term individually. First S_2,

$$S_2 = \mathbb{E}[|\text{tr}(x^T A x)|^3] \tag{292}$$

$$= \frac{1}{8}\mathbb{E}[|x^T H^T \Sigma^{-1} H x - x^T \mathbb{E}[H^T \Sigma^{-1} H]x|^3] \tag{293}$$

$$\leq \frac{1}{8}\mathbb{E}[|x^T H^T \Sigma^{-1} H x + x^T \mathbb{E}[H^T \Sigma^{-1} H]x|^3] \tag{294}$$

$$\leq \frac{1}{8}\mathbb{E}\left[\left|\frac{2n_t}{P}\|x\|_F^2\right|^3\right] \tag{295}$$

$$= \left(\frac{n_t}{P}\right)^3 \|x\|_F^6 \tag{296}$$

where

- (294) follows since $H^T \Sigma^{-1} H$ is PSD, and $\mathbb{E}[H^T \Sigma^{-1} H]$ is also PSD as a non-negative combination of PSD matrices, so that both $x^T H^T \Sigma^{-1} H x$ and $x^T \mathbb{E}[H^T \Sigma^{-1} H]x$ are non-negative
- (295) follows since $H^T \Sigma^{-1} H = V D V^T$ where

$$D = \text{diag}\left(c(\Lambda_1^2), \ldots, c(\Lambda_{n_{\min}}^2), 0, \ldots, 0\right) \tag{297}$$

and $D \leq \frac{n_t}{P} I_{n_t}$ in the PSD ordering, so

$$x^T H^T \Sigma^{-1} H x \leq \frac{n_t}{P} x^T V V^T x = \frac{n_t}{P}\|x\|_F^2 \tag{298}$$

and

$$x^T \mathbb{E}[H^T \Sigma^{-1} H]x \leq \frac{n_t}{P} x^T \mathbb{E}[V V^T]x = \frac{n_t}{P}\|x\|_F^2 \,. \tag{299}$$

Now we bound S_3 from (291),

$$S_3 = \mathbb{E}[|\text{tr}(x^T B)|^3] \tag{300}$$

$$= \mathbb{E}[\text{tr}(x^T H^T \Sigma^{-1} Z)|^3] \tag{301}$$

$$= \mathbb{E}\left[\left|\sum_{i=1}^{n_t}\sum_{j=1}^{T} \tilde{x}_{ij} Z_{ij} \frac{\Lambda_i}{1 + \frac{P}{n_t}\Lambda_i^2}\right|^3\right] \tag{302}$$

$$\leq n_t^2 T^2 \sum_{i=1}^{n_t}\sum_{j=1}^{T} \mathbb{E}\left[|\tilde{x}_{ij}|^3 |Z_{ij}|^3 \left|\frac{\Lambda_i}{1 + \frac{P}{n_t}\Lambda_i^2}\right|^3\right] \tag{303}$$

$$\leq \frac{n_t^2 T^2}{4}\left(\frac{n_t}{P}\right)^{3/2} \|x\|_F^3 \tag{304}$$

where

- In (302), define $\tilde{x} = V^T x$ and expand the trace.
- (303) follows from the triangle inequality, along with $|\sum_{i=1}^{n} a_i|^3 \leq n^2 \sum_{i=1}^{n} |a_i|^3$.
- (304) we have used $\mathbb{E}[|Z|^3] \leq 2$ for $Z \sim \mathcal{N}(0, 1)$ along with the bound

$$\left|\frac{x}{1 + ax^2}\right| \leq \frac{1}{2\sqrt{a}}\,. \tag{305}$$

Now notice that

$$\sum_{i=1}^{n_t}\sum_{j=1}^{T} |\tilde{x}_{ij}|^3 \leq \left(\sum_{i=1}^{n_t}\sum_{i=1}^{T} \tilde{x}_{ij}^2\right)^{3/2} \tag{306}$$

41

which can be viewed as the norm inequality $\|a\|_3 \leq \|a\|_2$ for $a \in \mathbb{R}^d$. Finally, we use $\|V^T x\|_F^2 = \|x\|_F^2$ for any orthogonal matrix V.

For the denominator in (138), the expression for $\frac{1}{T}\mathrm{Var}(W_j)$ is given in (55)-(59). Note that the final term (59) is non-negative, so we have the lower bound

$$\sum_{j=1}^{n} \mathrm{Var}(W_j) \geq K_1' n + K_2' \sum_{j=1}^{n} \left(\|x_j\|_F^2 - TP\right)^2 \tag{307}$$

$$\geq \max\left(nK_1', K_2' \sum_{j=1}^{n} \left(\|x_j\|_F^2 - TP\right)^2\right) \tag{308}$$

where

$$K_1' = T^2 \mathrm{Var}\left(C_r(H, P)\right) + T \sum_{i=1}^{n_{\min}} \mathbb{E}\left[V_{AWGN}\left(\frac{P}{n_t}\Lambda_i^2\right)\right] \tag{309}$$

$$K_2' = T\left(\frac{\|x\|_F^2}{n_t} - \frac{TP}{n_t}\right)^2 . \tag{310}$$

Hence $K_1' > 0$ whenever $P > 0$. Note that we use the assumption $\|x^n\|_F^2 = nTP$ freely here, as stated before. The lower bound on the variance (308), we obtain the upper bound

$$B_n(x^n) \leq \sqrt{n} \frac{\sum_{j=1}^{n} K_1 \|x_j\|_F^6 + K_2 \|x_j\|_F^3 + K_3}{\left(\max\left(nK_1', K_2' \sum_{j=1}^{n} \left(\|x_j\|_F^2 - TP\right)^2\right)\right)^{3/2}} \tag{311}$$

where all constant are non-negative. There are two cases based on which term achieves the max in the dominator. First, suppose

$$nK_1' \geq K_2' \sum_{j=1}^{n} \left(\|x_j\|_F^2 - TP\right)^2 . \tag{312}$$

Expanding the square yields

$$K_2' \sum_{j=1}^{n} \|x_j\|_F^4 \leq nK_1' + nT^2 P^2 K_2' . \tag{313}$$

Thus the terms in the numerator are bounded by

$$\sum_{j=1}^{n} \|x_i\|_F^6 \leq \left(\max_{i=1}^{n} \|x_i\|_F^2\right) \sum_{j=1}^{n} \|x_i\|_F^4 \leq n^{3/2}\delta^2(K_1' + T^2 P^2 K_2') \tag{314}$$

$$\sum_{j=1}^{n} \|x_i\|_F^3 \leq n^{1/4} \sum_{j=1}^{n} \|x_i\|_F^4 \leq n^{5/4}(K_1' + T^2 P^2 K_2') \tag{315}$$

where (314) uses the assumption $\|x_j\|_F \leq \delta n^{\frac{1}{4}}$. Applying this to B_n in (311), we see that in this case,

$$B_n(x^n) \leq \sqrt{n}\delta^2 C_1 + n^{1/4} C_2 + \frac{C_3}{n^{1/2}} \tag{316}$$

where the constant C_1, C_2, C_3 are non-negative constants.

Now take the case when

$$K_2' \sum_{j=1}^{n} \left(\|x_j\|_F^2 - TP\right)^2 \geq nK_1' . \tag{317}$$

42

Note that since $K_1' > 0$, in the case we must also have $K_2' > 0$ for the above inequality to hold. Let a be defined as follows

$$a = \frac{T^2 P^2}{T^2 P^2 + \frac{K_1'}{K_2'}} \ . \tag{318}$$

Here $a < 1$ since $K_1'/K_2' > 0$. Applying (317) yields

$$a \sum_{j=1}^n \|x_j\|_F^4 \geq a \left(n \frac{K_1'}{K_2'} + nT^2 P^2 \right) \tag{319}$$

$$\geq nT^2 P^2 \ . \tag{320}$$

With this, from (311) we obtain the following upper bound

$$B_n(x^n) \leq \sqrt{n} \frac{\sum_{j=1}^n K_1 \|x_j\|_F^6 + K_2 \|x_j\|_F^3 + K_3}{K_2'^{3/2} \left((1-a) \sum_{j=1}^n \|x_j\|_F^4 + a \sum_{j=1}^n \|x_j\|_F^4 - nT^2 P^2 \right)^{3/2}} \tag{321}$$

$$\leq \sqrt{n} \frac{\sum_{j=1}^n K_1 \|x_j\|_F^6 + K_2 \|x_j\|_F^3 + K_3}{K_2'^{3/2} \left((1-a) \sum_{j=1}^n \|x_j\|_F^4 \right)^{3/2}} \ . \tag{322}$$

where (322) uses (320). Now, we can upper bound each term in (322) as

$$\frac{K_1 \sum_{j=1}^n \|x_j\|_F^6}{K_2'^{3/2} \left((1-a) \sum_{j=1}^n \|x_j\|_F^4 \right)^{3/2}} \leq \frac{K_1 \max_{i=1,\dots,n} \|x_i\|_F^2}{K_2'^{3/2} (1-a)^{3/2} \left(\sum_{j=1}^n \|x_j\|_F^4 \right)^{1/2}} \tag{323}$$

$$\leq \frac{K_1 \delta^2 n^{1/2}}{n^{1/2} K_2'^{3/2} (1-a)^{3/2} (T^2 P^2 + nK_1')^{1/2}} \tag{324}$$

$$\frac{K_2 \sum_{j=1}^n \|x_j\|_F^3}{K_2'^{3/2} \left((1-a) \sum_{j=1}^n \|x_j\|_F^4 \right)^{3/2}} \leq \frac{K_2 n^{1/4}}{K_2'^{3/2} (1-a)^{3/2} \left(\sum_{j=1}^n \|x_j\|_F^4 \right)^{3/4}} \tag{325}$$

$$\leq \frac{K_2 n^{1/4}}{n^{1/2} K_2'^{3/2} (1-a)^{3/2} (T^2 P^2 + nK_1')^{3/4}} \tag{326}$$

$$\frac{K_3}{K_2'^{3/2} \left((1-a) \sum_{j=1}^n \|x_j\|_F^4 \right)^{3/2}} \leq \frac{K_3}{n^{3/2} \left(K_2' (1-a) (T^2 P^2 + nK_1') \right)^{3/2}} \tag{327}$$

where in (325) we have used $\sum_{i=1}^n a_i^3 \leq n^{1/4} \left(\sum_{i=1}^n a_i^4 \right)^{3/4}$ (easily obtained from p-norm inequalities), and both (323) and (327) use the assumption $\|x_j\|_F \leq \delta n^{\frac{1}{4}}$. Using these bounds in (322), we obtain

$$B_n(x^n) \leq \sqrt{n} \delta^2 C_1' + n^{1/4} C_2' + \frac{C_3'}{n^{1/2}} \tag{328}$$

where C_1', C_2', C_3' are non-negative constants.

From (316) and (328), we conclude that

$$B_n(x^n) \leq \sqrt{n} \delta^2 C_1'' + n^{1/4} C_2'' + \frac{C_3''}{n^{1/2}} \ . \tag{329}$$

$\square$

References

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] R. L. Dobrushin, "Mathematical problems in the Shannon theory of optimal coding of information," in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. 1, Berkeley, CA, USA, 1961, pp. 211–252.

[3] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," in *Trans. 3d Prague Conf. Inf. Theory*, Prague, 1962, pp. 689–723.

[4] Y. Polyanskiy and S. Verdú, "Finite blocklength methods in information theory (tutorial)," in *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/ISIT13_tutorial.pdf

[5] V. Y. Tan *et al.*, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Foundations and Trends® in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.

[6] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecom.*, vol. 10, no. 6, pp. 585–595, 1999.

[7] 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.

[8] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 8, pp. 1451–1458, 1998.

[9] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, 1999.

[10] Y. Polyanskiy and S. Verdú, "Scalar coherent fading channel: dispersion analysis," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Saint Petersburg, Russia, Aug. 2011, pp. 2959–2963.

[11] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Diversity versus channel knowledge at finite block-length," in *IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012, pp. 577–581.

[12] ——, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, July 2014.

[13] E. MolavianJazi and J. N. Laneman, "On the second-order coding rate of non-ergodic fading channels," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, USA, Oct. 2013.

[14] J. Hoydis, R. Couillet, and P. Piantanida, "The second-order coding rate of the MIMO Rayleigh block-fading channel," *IEEE Trans. Inf. Theory*, 2013, submitted. [Online]. Available: http://arxiv.org/abs/1303.3400

[15] S. Vituri, "Dispersion analysis of infinite constellations in ergodic fading channels," *arXiv preprint arXiv:1309.4638*, vol. M.S. Thesis, Tel Aviv University, 2013.

[16] A. Lapidoth, "On the high SNR capacity of stationary Gaussian fading channels," in *Proc. 2003 41st Allerton Conference*, vol. 41, no. 1, Allerton Retreat Center, Monticello, IL, USA, 2003, pp. 410–419.

[17] W. Yang, G. Durisi, and Y. Polyanskiy, "Minimum energy to send k bits over multiple-antenna fading channels," *arXiv preprint*, July 2015. [Online]. Available: http://arxiv.org/pdf/1507.03843v1.pdf

[18] W. Yang, G. Caire, G. Durisi, and Y. Polyanskiy, "Optimum power control at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4598 – 4615, Sept. 2015.

[19] W. Yang, A. Collins, G. Durisi, Y. Polyanskiy, and H. V. Poor, "A beta-beta achievability bound with applications," in *Proc. 2016 IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016.

[20] A. Collins and Y. Polyanskiy, "Orthogonal designs optimize achievable dispersion for coherent miso channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, Hawaii, July 2014, pp. 2524–2582.

[21] ——, "Dispersion of the coherent MIMO block-fading channel," in *Proc. 2016 IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1068–1072.

[22] Spectre, "SPECTRE: Short packet communication toolbox," 2015, GitHub repository. [Online]. Available: https://github.com/yp-mit/spectre

[23] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communication aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[24] J. Radon, "Lineare scharen orthogonaler matrizen," in *Abh. Sem. Hamburg*, vol. 1. Springer, 1922, pp. 1–14.

[25] Y. Polyanskiy and Y. Wu, *Lecture notes on Information Theory*, February 2016. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf

[26] S. Sandhu and A. Paulraj, "Space-time block codes: A capacity perspective," *IEEE Communications Letters*, vol. 4, no. 12, pp. 384–386, 2000.

[27] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, UK: Cambridge University Press, 2011.

[28] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010, available: http://www.princeton.edu/˜ypolyans.

[29] E. MolavianJazi and J. N. Laneman, "A second-order achievable rate region for Gaussian multi-access channels via a central limit theorem for functions," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6719–6733, 2015.

[30] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.

[31] X.-B. Liang, "Orthogonal designs with maximal rates," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2468–2503, 2003.

[32] G. Hamedani and M. Tata, "On the determination of the bivariate normal distribution from distributions of linear combinations of the variables," *American Mathematical Monthly*, pp. 913–915, 1975.

[33] M. F. Atiyah and I. G. Macdonald, *Introduction to commutative algebra*. Addison-Wesley Reading, 1969, vol. 2.