Communication on the Grassmann Manifold: A Geometric Approach to the Noncoherent Multiple-Antenna Channel

Lizhong Zheng, *Student Member, IEEE,* and David N. C. Tse, *Member, IEEE*

Abstract—In this paper, we study the capacity of multiple-antenna fading channels. We focus on the scenario where the fading coefficients vary quickly; thus an accurate estimation of the coefficients is generally not available to either the transmitter or the receiver. We use a noncoherent block fading model proposed by Marzetta and Hochwald. The model does not assume any channel side information at the receiver or at the transmitter, but assumes that the coefficients remain constant for a coherence interval of length T symbol periods. We compute the asymptotic capacity of this channel at high signal-to-noise ratio (SNR) in terms of the coherence time T, the number of transmit antennas M, and the number of receive antennas N. While the capacity gain of the coherent multiple antenna channel is $\min\{M, N\}$ bits per second per hertz for every 3-dB increase in SNR, the corresponding gain for the noncoherent channel turns out to be $M^*(1 - M^*/T)$ bits per second per herz, where $M^* = \min\{M, N, \lfloor T/2 \rfloor\}$. The capacity expression has a geometric interpretation as *sphere packing in the Grassmann manifold*.

Index Terms—Capacity, degees of freedom, multiple antennas, noncoherent communication, space–time coding.

I. INTRODUCTION

MOTIVATED by the need to increase the spectral efficiency of wireless systems, a major effort is being made to study the use of multiple antennas. While much work has been done on systems with multiple *receive* antennas, it was only recently shown by Foschini and Telatar [1]–[3] that much larger spectral efficiency can be achieved by utilizing multiple antennas at *both* the transmitter and the receiver.

In a single-antenna additive white Gaussian noise (AWGN) channel, it is well known that at high signal-to-noise ratio (SNR), 1-bit per second per hertz (b/s/Hz) capacity gain can be achieved with every 3-dB increase in SNR. In contrast, for a multiple antenna system with M transmit and N receive antennas and independent and identically distributed (i.i.d.) Rayleigh fading between all antenna pairs, the capacity gain is $\min\{M, N\}$ bits per second per hertz for every 3-dB SNR

increase [2]. The parameter $\min\{M, N\}$ is the number of degrees of freedom per second per hertz provided by the multiple antenna channel, and is a key measure of performance. This observation suggests the potential for very sizable improvement in spectral efficiency.

The result above is derived under the key assumption that the instantaneous fading coefficients are known to the receiver. Thus, this result can be viewed as a fundamental limit for *coherent* multiple-antenna communications. In a fixed wireless environment, the fading coefficients vary slowly, so the transmitter can periodically send pilot signals to allow the receiver to estimate the coefficients accurately. In mobile environments, however, the fading coefficients can change quite rapidly and the estimation of channel parameters becomes difficult, particularly in a system with a large number of antenna elements. In this case, there may not be enough time to estimate the parameters accurately enough. Also, the time one spends on sending pilot signals is not negligible, and the tradeoff between sending more pilot signals to estimate the channel more accurately and using more time to communicate to get more data through becomes an important factor affecting performance. In such situations, one may also be interested in exploring schemes that do not need explicit estimates of the fading coefficients. It is therefore of interest to understand the fundamental limits of *noncoherent* multiple-antenna communications.

A line of work was initiated by Marzetta and Hochwald [4], [5] to study the capacity of multiple-antenna channels when neither the receiver nor the transmitter knows the fading coefficients of the channel. They used a block fading channel model where the fading gains are i.i.d. Rayleigh distributed and remain constant for T symbol periods before changing to a new independent realization. Under this assumption, they reached the conclusion that further increasing the number of transmit antennas M beyond T cannot increase the capacity. They also characterized certain structure of the optimal input distribution, and computed explicitly the capacity of the one transmit and one receive antenna case at high SNR.

In this paper, we will use the same model to study the channel capacity for general values of M transmit and N receive antennas. We will focus on the high SNR regime, not only because it is more tractable than the general problem, but also because this is the regime where multiple antennas yield the most significant capacity increase from the additional spatial degrees of freedom provided. The high SNR capacity for the single-antenna case is obtained in [5] from first principles, by direct analysis of the integral involved in the relevant mutual information

functional. It seems difficult to generalize their technique to the multiple-antenna case. Instead, a geometric approach is adopted in this paper. By transforming the problem into a new coordinate system, the underlying geometry is described more naturally and the input optimization problem can be easily solved. Using this method, we get the following results.

1) Let $K = \min\{M, N\}$. In the case $T \geq K + N$, as SNR $\to \infty$, we show that the channel capacity (b/s/Hz) is given by

$$C(\mathsf{SNR}) = K\left(1 - \frac{K}{T}\right)\log_2 \mathsf{SNR} + c + o(1)$$

where c is an explicitly computed constant that depends only on M, N, and T, and $o(1)$ is a term that goes to zero at high SNR.[1] We specify the optimal input distribution that asymptotically achieves this capacity. For the case $T < K + N$, we characterize the rate that capacity increases with SNR. We conclude that in both cases, for each 3-dB SNR increase, the capacity gain is

$$M^*\left(1 - \frac{M^*}{T}\right)(\text{b/s/Hz})$$

with $M^* = \min\{M, N, \lfloor T/2 \rfloor\}$. *This is the number of degrees of freedom for noncoherent block fading multiple-antenna communications*

2) We show that at high SNR, the optimal strategy is to use only M^* of the M available antennas. In particular, having more transmit antennas than receive antennas does not provide any capacity increase at high SNR.

3) We show that given a coherence time T, the maximum number of degrees of freedom is achieved by using $T/2$ transmit antennas.

4) We give a geometric interpretation of the capacity expression as *sphere packing in the Grassmann manifold* $G(T, K)$: the set of all K-dimensional subspaces of $\mathcal{C}^T$.

5) We evaluate the performance of a scheme using training sequences and compare it with the capacity result. We show that it attains the full number of degrees of freedom.

At the end of the paper, we briefly contrast the high SNR regime with the low SNR regime, where the capacity of the multiple-antenna channel can be easily computed. We find that multiple antennas have a more significant impact in the high SNR regime than in the low SNR regime.

In this paper, the following notations will be used. We will use capital letters to indicate matrices, small letters for vectors and scalars, and boldfaced letters for random objects. For example, we write $\boldsymbol{X}$, $\boldsymbol{H}$ for random matrices, X, Y for deterministic matrices, θ, $\boldsymbol{x}$ for random vectors, and σ^2 for scalars. The only exception is SNR, which we use to denote the average signal-to-noise ratio at each receive antenna. Unless otherwise stated, we write $h(.)$ as differential entropy to the base e.

II. SYSTEM MODEL AND PRELIMINARIES

A. System Model

We follow the model in [5]. Assume the system has M transmit and N receive antennas, with i.i.d. Gaussian noise

at each of the receive antennas. The propagation coefficients form a $N \times M$ random matrix which neither the transmitter nor the receiver knows. We adopt a Rayleigh-fading model. We also assume that the coefficients remain constant for a time period T, and change to a new independent realization in the next time period. This can be a model for frequency hopping, ideally interleaved time division multiple access (TDMA) or packet-based system where each frame of data sees an independent realization of the channel but the channel is constant within each frame. The important feature of this model is that the channel remains constant only for a finite duration, so that there is inherent channel uncertainty at the receiver. In the following sections, we refer to T as the *coherence time* of the channel.

Because of the independence between the different coherence intervals, to calculate channel capacity it is sufficient to study one coherence interval, where each transmit antenna sends a T-dimensional vector, and each receive antenna receives a T-dimensional vector. In complex baseband representation, the system can be written as follows:

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W} \qquad (1)$$

where $\boldsymbol{X} \in \mathcal{C}^{M \times T}$, and the row vectors $\boldsymbol{x}_i \in \mathcal{C}^T$, $i = 1, \ldots, M$ correspond to the transmitted signal at the ith transmit antenna. Similarly, $\boldsymbol{Y} \in \mathcal{C}^{N \times T}$, and each row vector $\boldsymbol{y}_j \in \mathcal{C}^T$, $j = 1, \ldots, N$, is the received signal for the jth receive antenna.

The propagation gain from the jth transmit antenna to the ith receive antenna $\boldsymbol{h}_{ij}$, $i = 1 \cdots M$, $j = 1 \cdots N$ are i.i.d. complex Gaussian $\mathcal{CN}(0, 1)$ distributed with density

$$p(\boldsymbol{h}_{ij}) = \frac{1}{\pi}\exp[-|\boldsymbol{h}_{ij}|^2].$$

The additive noise $\boldsymbol{W} \in \mathcal{C}^{N \times T}$ has i.i.d. entries $\boldsymbol{w}_{nt} \sim \mathcal{CN}(0, \sigma^2)$. We normalize the equation to let the average transmit power at each transmit antenna in one symbol period be 1, so the power constraint can be written as

$$E\left[\sum_{i=1}^{M}\sum_{t=1}^{T}|\boldsymbol{x}_{it}|^2\right] = MT. \qquad (2)$$

We refer to the SNR as the average SNR at each receive antenna. Under the normalization above $\mathsf{SNR} = M/\sigma^2$.

The capacity (b/s/Hz) of the channel is given by

$$C_{M, N}(\mathsf{SNR}) = \frac{1}{T}\sup_{p_{\boldsymbol{x}}(\cdot)} I(\boldsymbol{X}; \boldsymbol{Y}) \qquad (3)$$

with the subscript indicating the number of antennas available. The optimization is over all input distributions of $\boldsymbol{X}$ satisfying the power constraint (2).

The goal of this paper is to compute high SNR approximations to $C_{M, N}(\mathsf{SNR})$ for various values of M, N, and T. All approximations are in the sense that the difference between the approximation and $C(\mathsf{SNR})$ goes to zero as the SNR tends to infinity.

B. Known Results

For the multiple-antenna channel with perfect knowledge of the fading coefficients at the receiver (but not at the transmitter), the channel capacity is computed in [1], [3]. We cite the main result in the following lemma.

---

[1]Since $\log \mathsf{SNR} \to \infty$ as $\mathsf{SNR} \to \infty$, this is a much more accurate approximation than, say, the statement that $\lim[C(\mathsf{SNR})/\log(\mathsf{SNR})] = K(1 - \frac{K}{T})$.

Lemma 1: Assume the fading coefficient matrix $\boldsymbol{H}$ is known to the receiver, the channel capacity (b/s/Hz) of a system with M transmit and N receive antennas is given by

$$
\begin{aligned}
C_{\text{coherent}}(\mathsf{SNR}) &= E\left[\log_2 \det\left(I_N + \frac{\mathsf{SNR}}{M}\boldsymbol{H}\boldsymbol{H}^\dagger\right)\right] \\
&= E\left[\log_2 \det\left(I_M + \frac{\mathsf{SNR}}{M}\boldsymbol{H}^\dagger\boldsymbol{H}\right)\right]. \quad (4)
\end{aligned}
$$

Defining $K = \min\{M, N\}$, $K' = \max\{M, N\}$, then a lower bound can be derived

$$
C_{\text{coherent}}(\mathsf{SNR}) \geq K \log_2 \frac{\mathsf{SNR}}{M} + \sum_{i=K'-K+1}^{K'} E[\log_2 \chi_{2i}^2]
$$

where χ_{2i}^2 is chi-square random variable with dimension $2i$. Moreover, this lower bound is asymptotically tight at high SNR. We observe that this is equivalent to the capacity of $K = \min\{M, N\}$ subchannels. In other words, the multiple-antenna channel has K *degrees of freedom* to communicate.

For the case $M = N$, at high SNR

$$
C_{\text{coherent}}(\mathsf{SNR}) = M \log_2 \frac{\mathsf{SNR}}{M} + \sum_{i=1}^{M} E[\log_2 \chi_{2i}^2] + o(1).
$$

If we let the number of antennas M increase to infinity, the high SNR capacity increases linearly with M, and

$$
\lim_{M \to \infty} \lim_{\mathsf{SNR} \to \infty} \left[\frac{C_{\text{coherent}}(\mathsf{SNR})}{M} - \log_2\left(\frac{\mathsf{SNR}}{e}\right)\right] = 0. \quad (5)
$$

This capacity can be achieved by using a "layered space–time architecture" which is discussed in detail in [1]. In the following, we will refer to this capacity result with the assumption of perfect knowledge of fading coefficients $\boldsymbol{H}$ as the *coherent capacity* of the multiple-antenna channel. In contrast, we use *noncoherent capacity* to denote the channel capacity with no prior knowledge of $\boldsymbol{H}$.

We now review several results for the noncoherent capacity from [4], [5].

Lemma 2: For any coherence time T and any number of receive antennas, the noncoherent capacity obtained with $M > T$ transmit antennas can also be obtained by $M = T$ transmit antennas.

As a consequence of this lemma, we will consider only the case of $M \leq T$ for the rest of the paper.

A partial characterization of the optimal input distribution is also given in [4]. Before presenting that result, we will first introduce the notion of *isotropically distributed* (i.d.) random matrices.

Definition 3: A random matrix $\boldsymbol{R} \in \mathcal{C}^{M \times T}$, for $T \geq M$, is called isotropically distributed (i.d.) if its distribution is invariant under rotation, i.e.,

$$
p(\boldsymbol{R}) = p(\boldsymbol{R}Q)
$$

for any deterministic $T \times T$ unitary matrix Q.

The following lemma gives an important property of i.d. matrices.

Lemma 4: If $\boldsymbol{H}$ is i.d., $\boldsymbol{Q}$ is a random unitary matrix that is independent of $\boldsymbol{H}$, then $\boldsymbol{H}\boldsymbol{Q}$ is independent of $\boldsymbol{Q}$.

To see this, observe that conditioning on any realization of $\boldsymbol{Q} = Q$, $\boldsymbol{H}Q$ has the same distribution as $\boldsymbol{H}$; thus, $\boldsymbol{H}\boldsymbol{Q}$ is independent of $\boldsymbol{Q}$.

Lemma 5: The input distribution that achieves capacity can be written as $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is an $M \times T$ i.d. unitary matrix, i.e., $\boldsymbol{\Theta}\boldsymbol{\Theta}^\dagger = I_M$. $\boldsymbol{A}$ is an $M \times M$ real diagonal matrix such that the joint distribution of the diagonal entries is exchangeable (i.e., invariant to the permutation of the entries). Moreover, $\boldsymbol{\Theta}$ and $\boldsymbol{A}$ are independent of each other.

The ith row θ_{x_i} of $\boldsymbol{\Theta}$ represents the direction of the transmitted signal from antenna i, i.e., $\theta_{x_i} = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|$. The ith diagonal entry of $\boldsymbol{A}$, $\boldsymbol{A}_{ii} = \|\boldsymbol{x}_i\|$, represents the norm of that signal. This characterization reduces the dimensionality of the optimization problem from MT to M by specifying the distribution of the signal directions, but the distribution of the norms is not specified. For the rest of the paper, we will, without loss of generality, consider input distributions within this class. The conjecture that constant equal power input $P(\boldsymbol{A} = \sqrt{T}I_M) = 1$ is asymptotically optimal at high SNR was made in [5]. In the rest of this paper, we will obtain the asymptotically optimal input distribution and give explicit expressions for the high SNR capacity. It turns out that the conjecture is true in certain cases but not in others.

C. Stiefel and Grassmann Manifolds

Natural geometric objects of relevance to the problem are the Stiefel and Grassmann manifolds. The *Stiefel manifold* $S(T, M)$ for $T \geq M$ is defined as the set of all unitary $M \times T$ matrices, i.e.,

$$
S(T, M) = \left\{Q \in \mathcal{C}^{M \times T} : QQ^\dagger = I_M\right\}.
$$

In the special case of $M = 1$, this is simply the surface of the unit sphere in $\mathcal{C}^T$.

The Stiefel manifold $S(T, M)$ can be viewed as an embedded submanifold of $\mathcal{C}^{M \times T}$ of real dimension $2TM - M^2$. One can define a measure μ on the Stiefel manifold, called the *Haar measure*, induced by the Lebesgue measure on $\Re^{2TM - M^2}$ through this embedding. It can be shown that this measure is invariant under rotation, i.e., if $\mathcal{S}$ is a measurable subset of $S(T, M)$, $\mu(\mathcal{S}) = \mu(\mathcal{S}P)$, for any unitary $T \times T$ matrix P. Hence, an i.d. unitary matrix is uniformly distributed on the Stiefel manifold with respect to the Haar measure. In the case $M = 1$, the Haar measure is simply the uniform measure on the surface of the unit sphere.

The total volume of the Stiefel manifold as computed from this measure is given by

$$
|S(T, M)| = \mu(S(T, M)) = \prod_{i=T-M+1}^{T} \frac{2\pi^i}{(i-1)!}. \quad (6)
$$

We can define the following equivalence relation on the Stiefel manifold: two elements $P, Q \in S(T, M)$ are equivalent if the row vectors (T-dimensional) span the same subspace, i.e., $P = UQ$ for some unitary $M \times M$ matrix U. The *Grassmann manifold* $G(T, M)$ is defined as the quotient space

of $S(T, M)$ with respect to this equivalence relation. Each element in the Grassmann manifold $G(T, M)$ is an equivalence class in $S(T, M)$. In other words, $G(T, M)$ is the set of all M-dimensional subspaces of $\mathcal{C}^T$.

For simplicity, in the rest of this paper, we will refer to "dimension" as complex dimension, where one complex dimension corresponds to two real dimensions. Thus, the dimensionality of the Grassmann manifold is given by

$$
\begin{aligned}
\dim(G(T, M)) &= \dim(S(T, M)) - \dim(S(M, M)) \\
&= M(T - M).
\end{aligned}
$$

The Haar measure on the Stiefel manifold induces a natural measure on the Grassmann manifold. The resulting volume of the Grassmann manifold is

$$
\begin{aligned}
|G(T, M)| &= \frac{|S(T, M)|}{|S(M, M)|} \\
&= \frac{\prod\limits_{i=T-M+1}^{T} \frac{2\pi^i}{(i-1)!}}{\prod\limits_{i=1}^{M} \frac{2\pi^i}{(i-1)!}}. \tag{7}
\end{aligned}
$$

For details concerning Stiefel manifolds, Grassmann manifolds, and the Haar measure, please refer to standard texts such as [6].

III. NONCOHERENT CAPACITY: $M = N$, $T \geq 2M$ CASE

In this section, we will study the multiple-antenna fading channel (1) with equal number of transmit and receive antennas, which will be referred as M throughout the section. We will first concentrate on the case that $T \geq 2M$. It turns out that this is the simplest case for which we can illustrate the use of a geometric approach. All other cases will be treated in Section IV.

To compute the channel capacity of the multiple-antenna channel, we need to compute the differential entropy of random matrices. To do this, a seemingly natural way is to view an $M \times T$ matrix as a vector of length MT, and compute the differential entropy in the rectangular coordinate system in $\mathcal{C}^{M \times T}$. However, the fact that the optimal input $\boldsymbol{X}$ has isotropic directions $\boldsymbol{\Theta}$ suggests the use of a different coordinate system. Therefore, we will start this section by introducing a new coordinate system. We will then transform the problem into this new coordinate system to calculate the relevant differential entropies and hence compute the channel capacity. A geometric interpretation of the result is given at the end of the section.

A. A New Coordinate System

An $M \times T$ matrix R, with $T \geq M$, can be represented as the subspace Ω_R spanned by its row vectors ($\Omega_R \in G(T, M)$), together with an $M \times M$ matrix C_R which specifies the M row vectors of R with respect to a canonical basis in Ω_R. The transformation

$$
R \to (C_R, \Omega_R) \tag{8}
$$

is a change of coordinate system $\mathcal{C}^{M \times T} \to \mathcal{C}^{M \times M} \times G(T, M)$. The Grassmann manifold $G(T, M)$ has $M(T - M)$ degrees of

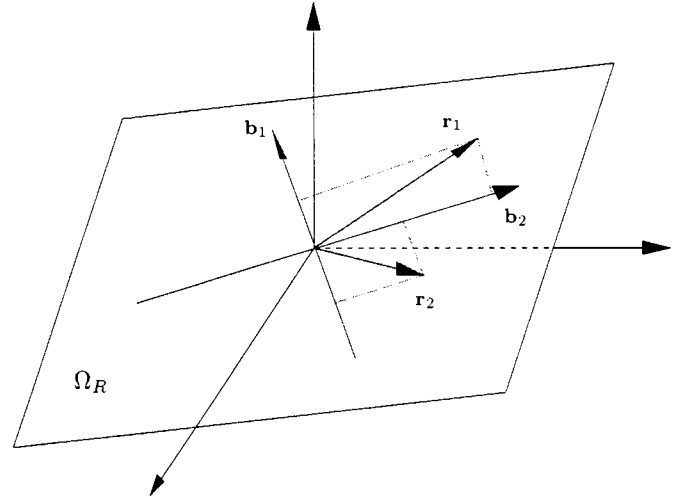Fig. 1.   Coordinate change in (8): $[b_1, \; b_2]$ is a basis of Ω_R, r_1, r_2 are the row vectors of R. $C_R = [c_{ij}]$ where c_{ij} is the length of the component of r_i in the direction of b_j.

freedom as discussed in Section II-C. This coordinate system is depicted in Fig. 1.

To understand the motivation of using such a coordinate system, we will first consider the channel without the additive noise $\boldsymbol{W}$: $Y_0 = \boldsymbol{HX}$. In this extreme case, the row vectors of the received signal $\boldsymbol{Y}_0$ span the same subspace as those of $\boldsymbol{X}$, i.e., $\Omega_{\boldsymbol{HX}} = \Omega_{\boldsymbol{X}}$, with probability 1. This fact shows that the random fading coefficients $\boldsymbol{H}$ affect the transmitted signals $\boldsymbol{X}$ by changing $\boldsymbol{C_X}$, but leave the subspace $\Omega_{\boldsymbol{X}}$ unchanged.

For the channel with additive noise, the subspace $\Omega_{\boldsymbol{X}}$ is corrupted only by the noise, but $\boldsymbol{C_X}$ is corrupted by both the noise and the channel fading. Essentially, the value of the coordinate system defined in (8) is to decompose $\mathcal{C}^{M \times T}$ into the directions that are affected by both the fading and the additive noise, and the directions that are affected by the additive noise alone. In the high SNR regime, the randomness of $\boldsymbol{C_X}$ is dominated by the randomness from the fading coefficients, rather than from the additive noise. Intuitively, we can think that $\boldsymbol{C_X}$ is corrupted only by the channel fading. Thus, the use of coordinate system (8) allows us to consider the effect of the fading and the additive noise separately at high SNR.

The following lemma provides a connection between the differential entropies computed in rectangular coordinates and in the new coordinate system.

Lemma 6 (Change of Coordinates): Let $\boldsymbol{R} \in \mathcal{C}^{M \times T}$ be a random matrix, $T \geq M$. If $\boldsymbol{R}$ is i.d., i.e.,

$$
p(\boldsymbol{R}) = p(\boldsymbol{R}Q), \qquad \forall \text{ deterministic unitary matrix } Q \in \mathcal{C}^{T \times T} \tag{9}
$$

then

$$
h(\boldsymbol{R}) = h(\boldsymbol{C_R}) + \log |G(T, M)| + (T - M)E[\log \det \boldsymbol{R}\boldsymbol{R}^\dagger] \tag{10}
$$

where $|G(T, M)|$ is given by (7).

Remarks: Notice that the differential entropies $h(.)$ in (10) are computed in different coordinate systems. $h(\boldsymbol{R})$ is computed

in the rectangular coordinates in $\mathcal{C}^{M \times T}$, and $h(\boldsymbol{C_R})$ in $\mathcal{C}^{M \times M}$. In the rest of the paper, we write $h(.)$ of a random matrix without detailed explanation on the coordinate systems. If the argument has certain properties (e.g., diagonal, unitary, triangular), the entropy is calculated in the corresponding subspace instead of the whole space.

The term $h(\boldsymbol{C_R}) + \log |G(T, M)|$ in the right-hand side of (10) can be interpreted as the differential entropy of $\boldsymbol{R}$ computed in $\mathcal{C}^{M \times M} \times G(T, M)$. For a general matrix R, C_R depends on the choice of the canonical basis of Ω_R. For each choice of a basis, (8) gives a different coordinate change. However, with the additional assumption (9), the distribution of $\boldsymbol{C_R}$ does not depend on the choice of basis. To see this, we first factorize $\boldsymbol{R}$ via the LQ decomposition

$$\boldsymbol{R} = \boldsymbol{L V}^\dagger \qquad (11)$$

where $\boldsymbol{L} \in \mathcal{C}^{M \times M}$ is lower triangular with real nonnegative diagonal entries. $\boldsymbol{V} \in \mathcal{C}^{T \times M}$ is a unitary matrix. Now the assumption (9) is equivalent to

$$\boldsymbol{V} \text{ is i.d. and independent of } \boldsymbol{L}. \qquad (12)$$

Under this assumption, the row vectors of $\boldsymbol{R}$ are i.d. in $\mathcal{C}^T$, which implies that the subspace spanned by these row vectors $\Omega_{\boldsymbol{R}}$ is uniformly distributed in the Grassmann manifold $G(T, M)$. Furthermore, given $\Omega_{\boldsymbol{R}}$, the row vectors are i.d. in $\Omega_{\boldsymbol{R}}$. Therefore, irrespective of the basis chosen, the coefficient matrix $\boldsymbol{C_R}$ has the same distribution as $\boldsymbol{LQ}$, for an i.d. unitary matrix $\boldsymbol{Q} \in \mathcal{C}^{M \times M}$ that is independent of $\boldsymbol{L}$.

It is well known that for the same random object, the differential entropies computed in different coordinate systems differ by $E[\log \boldsymbol{J}]$, where $\boldsymbol{J}$ is the Jacobian of the coordinate change. The term $(T - M)E[\log \det \boldsymbol{R R}^\dagger]$ in (10) is, in fact, the Jacobian term for the coordinate change (8). To prove that and to prove Lemma 6, we need to first study the Jacobian of some standard matrix factorizations. It is a well-established approach in multivariate statistical analysis to view matrix factorizations as changing of coordinate systems. For example, the LQ decomposition (11) can be viewed as a coordinate change $\mathcal{C}^{M \times T} \to \mathcal{L} \times S(T, M)$, where $\mathcal{L}$ is the set of all lower triangular matrices with real nonnegative diagonal entries. A brief introduction of this technique is given in Appendix A. The Jacobian of the LQ coordinate change is given in the following lemma.

Lemma 7 [7]: Let $l_{11}, \ldots, l_{MM}$ be the diagonal elements of L. The Jacobian of the LQ decomposition (11) is

$$J_{T, M} = \prod_{i=1}^{M} l_{ii}^{2(T-i)}. \qquad (13)$$

Proof of Lemma 6: We observe that the coordinate change (8) can be obtained by consecutive uses of the LQ decomposition as follows: by Lemma 7 and (12)

$$h(\boldsymbol{R}) = h(\boldsymbol{L}) + h(\boldsymbol{V}^\dagger) + E[\log J_{T, M}]$$
$$= h(\boldsymbol{L}) + \log |S(T, M)| + E[\log J_{T, M}]$$

and

$$h(\boldsymbol{C_R}) = h(\boldsymbol{LQ})$$
$$= h(\boldsymbol{L}) + \log |S(M, M)| + E[\log J_{M, M}].$$

Combine the two equations and we get

$$h(\boldsymbol{R}) = h(\boldsymbol{C_R}) + \log |G(T, M)| + E\left[\log \prod_{i=1}^{M} l_{ii}^{2(T-M)}\right]$$
$$= h(\boldsymbol{C_R}) + \log |G(T, M)| + (T-M)E[\log \det \boldsymbol{R R}^\dagger]. \quad \square$$

B. Channel Capacity

For convenience, we will rewrite the channel model here

$$\boldsymbol{Y} = \boldsymbol{H X} + \boldsymbol{W} \qquad (14)$$

where $\boldsymbol{H} \in \mathcal{C}^{M \times M}$ is the matrix of fading coefficients with i.i.d. $\mathcal{CN}(0, 1)$ entries. $\boldsymbol{W} \in \mathcal{C}^{M \times T}$ is the additive Gaussian noise with i.i.d. $\mathcal{CN}(0, \sigma^2)$ entries. The input $\boldsymbol{X} \in \mathcal{C}^{M \times T}$ can be written as $\boldsymbol{X} = \boldsymbol{A\Theta}$, where $\boldsymbol{A} = \text{diag}(\|\boldsymbol{x}_i\|, i = 1, \ldots M)$, contains the norms of the transmitted vectors at each transmit antenna; $\boldsymbol{\Theta}$ is an i.d. unitary matrix, which is independent of $\boldsymbol{A}$. The total transmit power is normalized to be

$$\sum_{i=1}^{M} E[\|\boldsymbol{x}_i\|^2] \le MT$$

and the SNR is $\text{SNR} = M/\sigma^2$.

In this section, we will compute the mutual information $I(\boldsymbol{X}; \boldsymbol{Y})$ in terms of the input distribution of $\boldsymbol{A}$, and find the optimal input distribution to maximize the mutual information.

Now

$$I(\boldsymbol{X}; \boldsymbol{Y}) = h(\boldsymbol{Y}) - h(\boldsymbol{Y}|\boldsymbol{X}).$$

To compute $h(\boldsymbol{Y}|\boldsymbol{X})$, we observe that given $\boldsymbol{X}$, $\boldsymbol{Y}$ is Gaussian. The row vectors of $\boldsymbol{Y}$ are independent of each other, and have the common covariance matrix

$$K_{\boldsymbol{y}_i} = \boldsymbol{X}^\dagger \boldsymbol{X} + \sigma^2 I_T = \boldsymbol{\Theta}^\dagger \boldsymbol{A}^2 \boldsymbol{\Theta} + \sigma^2 I_T.$$

Therefore, the conditional entropy $h(\boldsymbol{Y}|\boldsymbol{X})$ is given by

$$h(\boldsymbol{Y}|\boldsymbol{X}) = ME\left[\sum_{i=1}^{M} \log \pi e(\|\boldsymbol{x}_i\|^2 + \sigma^2)\right]$$
$$+ M(T - M) \log \pi e \sigma^2. \quad (15)$$

Now since we only need to compute $h(\boldsymbol{Y})$ for the optimal input distribution of $\boldsymbol{X}$, we will first characterize the optimal input distribution in the following lemma.

Lemma 8: Let $(\boldsymbol{x}_i^{(\sigma)}, i = 1, \ldots M)$ be the optimal input signal of each antenna at noise level σ^2. If $T \ge 2M$

$$\frac{\left\|\boldsymbol{x}_i^{(\sigma)}\right\|}{\sigma} \xrightarrow{P} \infty, \qquad \text{for } i = 1, \ldots M \qquad (16)$$

where $\xrightarrow{P}$ denotes convergence in probability as $\sigma^2 \to 0$.

Proof: See Appendix B. $\square$

This lemma says that to achieve the capacity at high SNR, the norm of the signal transmitted at each antenna must be much higher than the noise level. Essentially, this is similar to the situation that in the high SNR regime of the AWGN channel, it is much more preferable to spread the available energy over all degrees of freedom rather than transmit over only a fraction of the degrees of freedom.

Before using Lemma 8 to compute the channel capacity rigorously, we will first make a few approximations at high SNR to illustrate the intuition behind the complete calculation of the

capacity. We first observe that since $\|\boldsymbol{x}_i\|^2 \gg \sigma^2$ for all $i = 1, \ldots, M$

$$
\begin{aligned}
h(\boldsymbol{Y}|\boldsymbol{X}) &= ME\left[\sum_{i=1}^{M} \log \pi e(\|\boldsymbol{x}_i\|^2 + \sigma^2)\right] \\
&\quad + M(T-M)\log \pi e \sigma^2 \\
&\approx ME\left[\sum_{i=1}^{M} \log \pi e\|\boldsymbol{x}_i\|^2\right] + M(T-M)\log \pi e \sigma^2 \\
&= ME\left[\log \det \boldsymbol{A}^2\right] + M^2 \log \pi e \\
&\quad + M(T-M)\log \pi e \sigma^2.
\end{aligned} \tag{17}
$$

To compute $h(\boldsymbol{Y})$, we make the approximation

$$
h(\boldsymbol{Y}) \approx h(\boldsymbol{HX}).
$$

Now observe that $\boldsymbol{HX}$ is i.d., so we can apply Lemma 6. Notice that given Ω_X, $\boldsymbol{HX}$ is i.d. in the subspace; thus, $\boldsymbol{C_{HX}}$ has the same distribution as $\boldsymbol{HAQ}$, where $\boldsymbol{Q} \in \mathcal{C}^{M \times M}$ is i.d. unitary and is independent of $\boldsymbol{HA}$

$$
\begin{aligned}
h(\boldsymbol{HX}) &= h(\boldsymbol{C_{HX}}) + \log|G(T, M)| \\
&\quad + (T-M)E[\log \det(\boldsymbol{HA}^2\boldsymbol{H}^\dagger)] \\
&= h(\boldsymbol{HAQ}) + \log|G(T, M)| \\
&\quad + (T-M)E[\log \det \boldsymbol{A}^2] \\
&\quad + (T-M)E[\log \det \boldsymbol{HH}^\dagger].
\end{aligned} \tag{18}
$$

Combining (17) and (18), we have

$$
\begin{aligned}
I(\boldsymbol{X}; \boldsymbol{Y}) &\approx \log|G(T, M)| + (T-M)E[\log \det \boldsymbol{HH}^\dagger] \\
&\quad - M(T-M)\log \pi e \sigma^2 - M^2 \log \pi e \\
&\quad + h(\boldsymbol{HAQ}) + (T-2M)E[\log \det \boldsymbol{A}^2].
\end{aligned} \tag{19}
$$

Now observe that random matrix $\boldsymbol{HAQ} \in \mathcal{C}^{M \times M}$ has bounded total average power

$$
E\left[\sum_{i,j=1}^{M} |(\boldsymbol{HAQ})_{ij}|^2\right] = ME\left[\sum_{i=1}^{M} \|\boldsymbol{x}_i\|^2\right] \le M^2 T.
$$

Therefore, the differential entropy is maximized by the matrix with i.i.d. $\mathcal{CN}(0, T)$ entries, i.e., $h(\boldsymbol{HAQ}) \le M^2 \log \pi e T$. The equality is achieved by setting $\|\boldsymbol{x}_i\|^2 = T$ with probability 1 for all is. Since $T \ge 2M$, $(T-2M)E[\log \det \boldsymbol{A}^2]$ is also maximized by the same choice of input distribution, by the concavity of the log function. Thus, the equal constant norm input distribution maximizes the approximate mutual information, and the maximum value is

$$
\begin{aligned}
&\log|G(T, M)| + (T-M)E[\log \det \boldsymbol{HH}^\dagger] \\
&\qquad\qquad\qquad\qquad - M(T-M)\log(\pi e \sigma^2 / T).
\end{aligned}
$$

A precise statement of the result is contained in the following theorem.

Theorem 9: For the multiple-antenna channel with M transmit, M receive antennas, and coherence time $T \ge 2M$, the high SNR capacity (b/s/Hz) is given by

$$
C_{M,M}(\mathsf{SNR}) = M\left(1 - \frac{M}{T}\right)\log_2 \mathsf{SNR} + c_{M,M} + o(1) \tag{20}
$$

where

$$
\begin{aligned}
c_{M,M} &= \frac{1}{T}\log_2|G(T, M)| + M\left(1 - \frac{M}{T}\right)\log_2 \frac{T}{M\pi e} \\
&\quad + \left(1 - \frac{M}{T}\right)E[\log_2 \det \boldsymbol{HH}^\dagger]
\end{aligned}
$$

and

$$
E[\log \det \boldsymbol{HH}^\dagger] = \sum_{i=1}^{M} E[\log \chi_{2i}^2]
$$

with χ_{2i}^2 a Chi-square random variable of dimension $2i$.

Proof: See Appendix C. $\square$

To connect this result to the capacity of the coherent channel, we rewrite (20) as

$$
\begin{aligned}
C_{M,M}(\mathsf{SNR}) &= \left(1 - \frac{M}{T}\right)C_{\text{coherent}}(\mathsf{SNR}) + \frac{1}{T}\log_2|G(T, M)| \\
&\quad + M\left(1 - \frac{M}{T}\right)\log_2 \frac{T}{\pi e} + o(1)
\end{aligned} \tag{21}
$$

where $C_{\text{coherent}}(\mathsf{SNR})$ is the channel capacity with perfect knowledge of the fading coefficients, given in (4).

An important observation on the capacity result is that for each 3-dB SNR increase, the capacity gain is $M(1 - \frac{M}{T})$ (bits per second per hertz), the number of degrees of freedom in the channel.

If we fix the number of antennas M and let the coherence time T increase to infinity, this corresponds to the case with perfect knowledge of fading coefficients. Indeed, the capacity given in (21) converges to C_{coherent} as $T \to \infty$. To see this, we use Stirling's formula $n! \approx n^n e^{-n}\sqrt{2\pi n}$, and write

$$
\begin{aligned}
&\frac{1}{T}\log|G(T, M)| + M\left(1 - \frac{M}{T}\right)\log\frac{T}{\pi e} \\
&= \frac{1}{T}\left(\sum_{i=T-M+1}^{T} \log\frac{2\pi^i}{(i-1)!} - \sum_{i=1}^{M} \log\frac{2\pi^i}{(i-1)!}\right) \\
&\quad + M\left(1 - \frac{M}{T}\right)\log\frac{T}{\pi e} \\
&= \frac{M}{T}\log\frac{\pi^T}{T!} - \underbrace{\frac{1}{T}\sum_{i=1}^{M}\log\frac{\pi^{2i-1}T!}{(i-1)!(T-i+1)!}}_{\to 0} \\
&\quad + M\left(1 - \frac{M}{T}\right)\log\frac{T}{\pi e} \\
&\to \frac{M}{T}\log\frac{\pi^T}{T^T e^{-T}\sqrt{2\pi T}} + M\left(1 - \frac{M}{T}\right)\log\frac{T}{\pi e} \\
&\to \frac{M}{T}\log\frac{\pi^T e^T}{T^T} + M\log\frac{T}{\pi e} = 0.
\end{aligned}
$$

In Fig. 2, we plot the high SNR approximation of the non-coherent capacity given in (21), in comparison to the capacity with perfect knowledge C_{coherent}. We observe that as $T \to \infty$, the capacity given in (21) approaches $C_{\text{coherent}}(\mathsf{SNR})$.

In Fig. 3, we plot the high SNR noncoherent capacity for an 8 by 8 multiple-antenna channel in comparison to the single-antenna AWGN channel capacity with the same SNR. We observe that multiple antennas do provide a remarkable capacity gain even when the channel is not known at the receiver. This gain is a good fraction of the gain obtained when the channel is known.
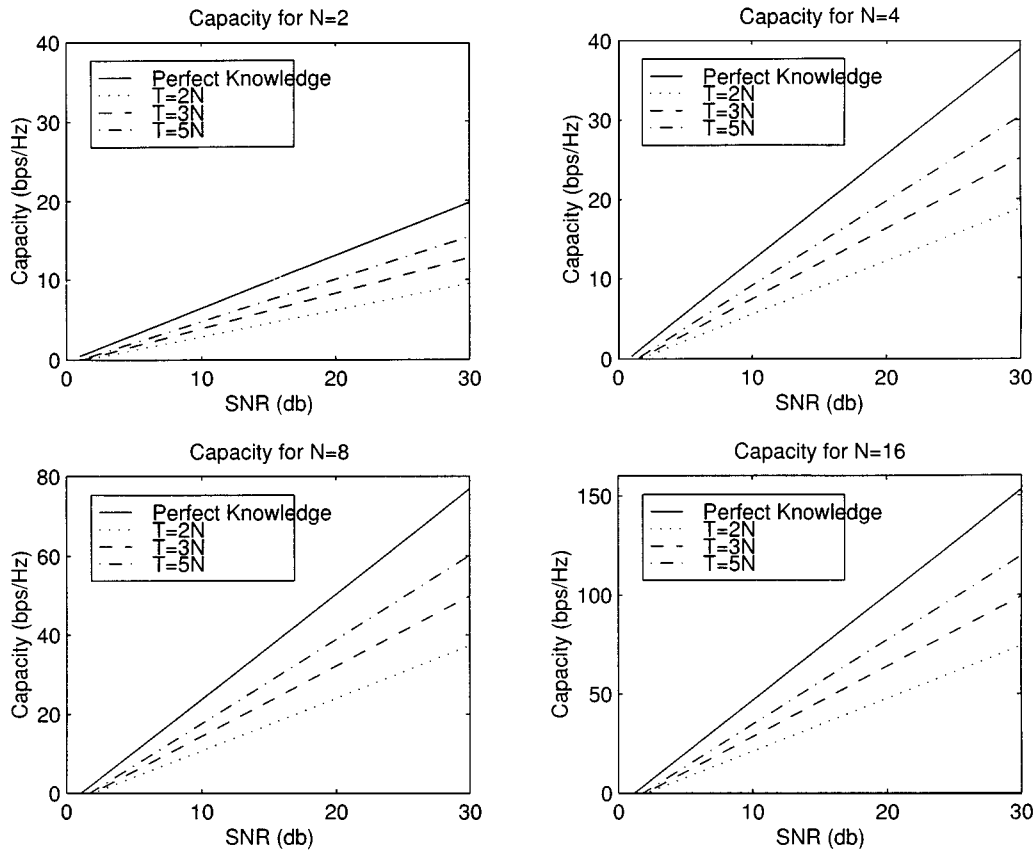
Fig. 2. Noncoherent channel capacity (high SNR approximation).

Corollary 10: For the special case $M = 1$, $T \geq 2$, the capacity (b/s/Hz) is

$$C_{1,1}(\mathsf{SNR}) = \frac{T-1}{T} C_{\text{coherent}}(\mathsf{SNR})$$
$$+ \frac{1}{T} \log_2 \left[\left(\frac{T}{e} \right)^{T-1} \frac{1}{(T-1)!} \right] + o(1).$$

This result is derived in [5] from first principles.

In the following corollary, we discuss the large system limit, where both M and T increase to infinity, with the ratio M/T fixed. As in the perfect knowledge case, the channel capacity increases linearly with the number of antennas M, when both M and SNR are large.

Corollary 11: For the case when both M and T approach infinity, but the ratio $\gamma = M/T$ is fixed, the channel capacity $C_{M,M}$ increases linearly with the number of antennas M. The ratio $C_{M,M}/M$ (b/s/Hz/transmit antenna) is given by

$$\lim_{M \to \infty} \lim_{\mathsf{SNR} \to \infty} \left[\frac{C_{M,M}(\mathsf{SNR})}{M} - \left(\frac{k(\gamma)}{\log_2 e} + (1-\gamma) \log_2 \frac{\mathsf{SNR}}{e} \right) \right] = 0 \quad (22)$$

where

$$k(\gamma) = \frac{(1-\gamma)^2}{2\gamma} \log(1-\gamma) + \frac{\gamma}{2} \log \gamma + \frac{1-\gamma}{2}.$$

Notice that the term $\log \mathsf{SNR}/e$ is the limiting coherent capacity per antenna C_{coherent}/M given in (5). It can be easily checked that $k(\gamma) < 0$ for all γ. This fact shows that to communicate in

noncoherent channel, we have to pay the price of M^2/T degrees of freedom, as well as an extra penalty of $k(\gamma)$ per antenna.

Proof: Consider

$$\frac{1}{M} C_{M,M}(\mathsf{SNR}) = \frac{1}{MT} \log |G(T, M)| + (1-\gamma) \log \frac{T}{\pi e}$$
$$+ (1-\gamma) \underbrace{\frac{1}{M} C_{\text{coherent}}(\mathsf{SNR})}_{\to \log \mathsf{SNR}/e \text{ by (5)}} + o(1).$$

Using the definition of $|G(T, M)|$ given in (7), the first term becomes

$$\frac{1}{MT} \log |G(T, M)|$$
$$= \frac{1}{MT} \left[\sum_{i=T-M+1}^{T} \frac{2\pi^i}{(i-1)!} - \sum_{i=1}^{M} \frac{2\pi^i}{(i-1)!} \right].$$

Now use Stirling's formula $n! \approx n^n e^{-n} \sqrt{2\pi n}$, and let M and T grow, we have

$$\frac{1}{MT} \log |G(T, M)|$$
$$\to \frac{1}{MT} \left[\sum_{i=T-M+1}^{T} i \log \frac{\pi e}{i} - \sum_{i=1}^{M} i \log \frac{\pi e}{i} \right]$$
$$= \frac{1}{M} \left[\sum_{i=T-M+1}^{T} \frac{i}{T} \log \frac{\pi e}{i/T} - \sum_{i=1}^{M} \frac{i}{T} \log \frac{\pi e}{i/T} \right]$$
$$+ (1-\gamma) \log \frac{1}{T}$$
$$\to \frac{1}{\gamma} \left[\int_{1-\gamma}^{1} t \log \frac{\pi e}{t} \, dt - \int_{0}^{\gamma} t \log \frac{\pi e}{t} \, dt \right] - (1-\gamma) \log T$$
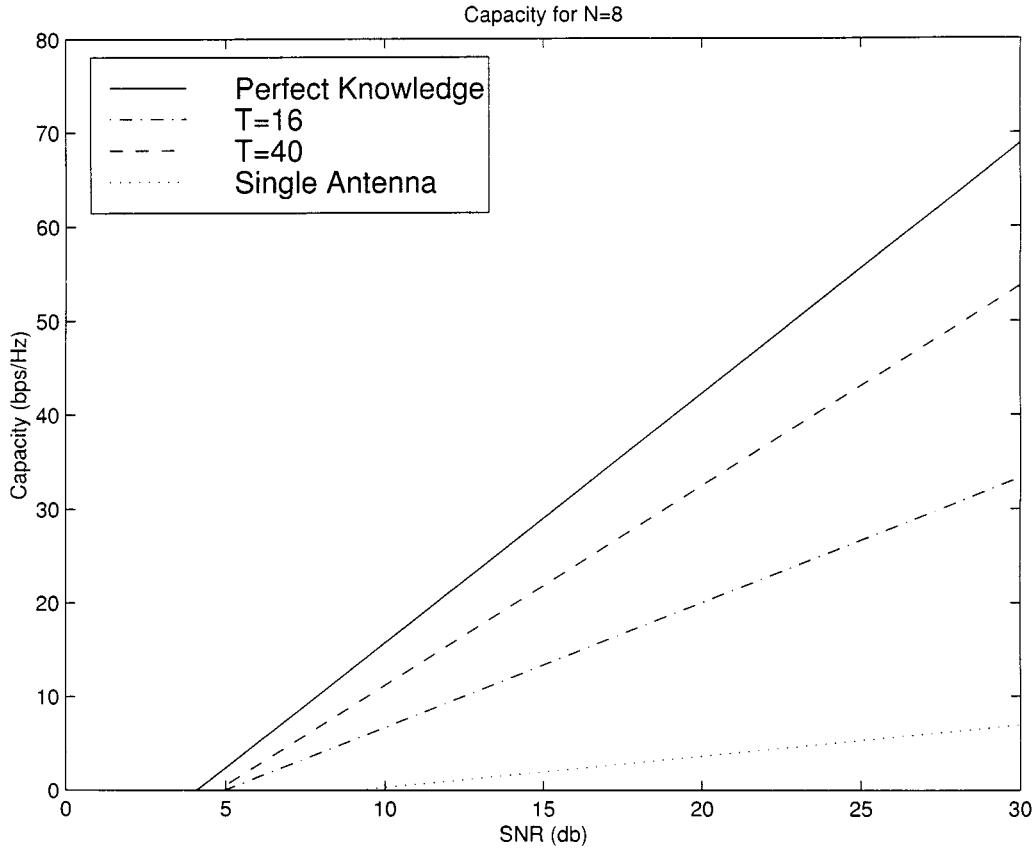
Fig. 3.   Comparison of noncoherent channel capacity versus AWGN capacity.

$$= (1 - \gamma) \log \frac{\pi e}{T} + \frac{1}{\gamma} \left[\int_0^\gamma t \log t \, dt - \int_{1-\gamma}^1 t \log t \, dt \right].$$

Hence,

$$\frac{1}{MT} \log |G(T, M)| + (1 - \gamma) \log \frac{T}{\pi e}$$
$$\to \frac{1}{\gamma} \left[\int_0^\gamma t \log t \, dt - \int_{1-\gamma}^1 t \log t \, dt \right]$$
$$= \frac{(1 - \gamma)^2}{2\gamma} \log(1 - \gamma) + \frac{\gamma}{2} \log \gamma + \frac{1 - \gamma}{2}. \qquad \square$$

C. Geometric Interpretation

By using the coordinate system (8), we can decompose the mutual information into two terms

$$I(\boldsymbol{X}; \boldsymbol{Y}) = I(\Omega_{\boldsymbol{X}}; \boldsymbol{Y}) + I(\boldsymbol{C}_{\boldsymbol{X}}; \boldsymbol{Y} | \Omega_{\boldsymbol{X}}). \qquad (23)$$

That is, we decompose the total mutual information into the mutual information conveyed by the subspace $\Omega_{\boldsymbol{X}}$, and the mutual information conveyed within the subspace.

Since $\boldsymbol{X}$ is of the form $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{\Theta}$, with $\boldsymbol{\Theta}$ being an i.d. unitary matrix independent of $\boldsymbol{A}$, we have $\boldsymbol{C}_{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is an i.d. $M \times M$ unitary matrix independent of $\boldsymbol{A}$. Consequently, we can write $\boldsymbol{C}_{\boldsymbol{H}\boldsymbol{X}} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{Q}$. From the previous section, we know that the asymptotically optimal input distribution at high SNR is the equal constant norm input

$$P(\|\boldsymbol{x}_i\| = \sqrt{T}) = 1, \qquad \forall i = 1, \dots, M.$$

With this input, $\boldsymbol{C}_{\boldsymbol{X}} = \sqrt{T}\boldsymbol{Q}$ and $\boldsymbol{C}_{\boldsymbol{H}\boldsymbol{X}} = \sqrt{T}\boldsymbol{H}\boldsymbol{Q}$. Observe that $\boldsymbol{H}$ is itself i.d., and by Lemma 4, $\boldsymbol{H}\boldsymbol{Q}$ is independent of $\boldsymbol{Q}$.

Therefore, $\boldsymbol{Y}$ is independent of $\boldsymbol{C}_{\boldsymbol{X}} = \sqrt{T}\boldsymbol{Q}$, i.e., the observation of $\boldsymbol{Y}$ provides no information about $\boldsymbol{C}_{\boldsymbol{X}}$; thus, the second term in (23) is 0. Now we conclude that by using the equal constant norm input, all the mutual information is conveyed by the random subspace $\Omega_{\boldsymbol{X}}$

$$I(\boldsymbol{X}; \boldsymbol{Y}) = I(\Omega_{\boldsymbol{X}}; \boldsymbol{Y}).$$

In the noncoherent multiple-antenna channel, the information-carrying object is a random subspace $\Omega_{\boldsymbol{X}}$, which is a random point in the Grassmann manifold. In contrast, for the *coherent* case, the information-carrying object is the matrix $\boldsymbol{X}$ itself. Thus, the number of degrees of freedom reduces from MT, the dimension of the set of M by T matrices in the coherent case, to $M(T - M)$, the dimension of the set of all *row spaces* of M by T matrices in the noncoherent case. The loss of M^2 degrees of freedom stems from the *channel uncertainty* at the receiver: unitary $M \times T$ matrices with the same row space cannot be distinguished at the receiver.

In the following, we will further discuss the capacity result to show that it has a natural interpretation as *sphere packing in the Grassmann manifold*.

In the canonical AWGN channel, the channel capacity has a well-known interpretation in terms of "sphere packing." This intuition can be generalized to coherent and noncoherent multiple-antenna channels.

For the coherent multiple-antenna channel, the high SNR channel capacity is given by $C \approx \log \det(\frac{\text{SNR}}{M} HH^\dagger)$. After appropriate scaling, we have the transmit power $E[\|\boldsymbol{x}_i\|^2] = 1$, and the noise variance $\sigma^2 = M/\text{SNR}$. Let the input $\boldsymbol{x}$ be i.i.d.
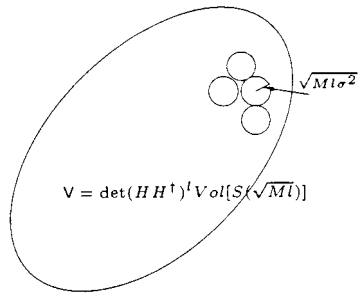
Fig. 4. Sphere packing in coherent multiple-antenna channel.

Gaussian distributed, and the codeword length be l. We denote $B_d(r)$ as the sphere of radius r in $\mathcal{C}^d$. For large l, the input sequence $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l)$ lies in the sphere $B_{Ml}(\sqrt{Ml})$ with high probability. The fading matrix H stretches $\boldsymbol{x}$ to $H\boldsymbol{x}$, which lies in an ellipsoid of volume $\det(HH^\dagger)^l Vol(B_{Ml}(\sqrt{Ml}))$. The received signal lies in a sphere $B_{Ml}(\sqrt{Ml\sigma^2})$ around $(H\boldsymbol{x}_1, \ldots, H\boldsymbol{x}_l)$. The capacity can be written as the logarithm of the ratio of the two volumes

$$C \approx \frac{1}{l} \log \frac{\det(HH^\dagger)^l Vol(B_{Ml}(\sqrt{Ml}))}{Vol(B_{Ml}(\sqrt{Ml\sigma^2}))}$$

$$= \log\det\left(\frac{1}{\sigma^2} HH^\dagger\right) = \log\det\frac{\mathsf{SNR}}{M} HH^\dagger.$$

The sphere packing is depicted in Fig. 4.

For the noncoherent channel where the fading coefficients are unknown, we can interpret the capacity by *sphere packing in the Grassmann manifold*. Since the subspace $\Omega_{\boldsymbol{X}}$ is the object that we use to convey information, we view the transmitted signal in each coherence interval as a point in the Grassmann manifold $G(T, M)$. Similar to the perfect knowledge case, $\boldsymbol{H}$ scales the volume to be $\det(T\boldsymbol{H}\boldsymbol{H}^\dagger)^{T-M}|G(T, M)|$. With codewords of length l, the received signal lies in the product space of l scaled Grassmann manifolds, with dimension $M(T-M)l$. The noise perturbs the signal in the sphere $B_{M(T-M)l}(\sqrt{M(T-M)l\sigma^2})$. Denote $H_i, i = 1, \ldots l$ as the fading coefficient matrix in coherence interval i, we write the ratio of the two volumes

$$q = \frac{\prod\limits_{i=1}^{l} \det(T\boldsymbol{H}_i\boldsymbol{H}_i^\dagger)^{T-M}|G(T, M)|}{Vol(B_{M(T-M)l}(\sqrt{M(T-M)l\sigma^2}))}$$

and

$$\frac{1}{l}\log q = (T-M)\frac{1}{l}\sum_{i=1}^{l}\log\det(T\boldsymbol{H}_i\boldsymbol{H}_i^\dagger) + \log|G(T, M)|$$
$$- \frac{1}{l}\log Vol\left(B_{M(T-M)l}\left(\sqrt{M(T-M)l\sigma^2}\right)\right).$$

Using the formula $Vol(B_n(r)) = \pi^n r^{2n}/n!$ and Stirling's formula $n! \approx n^n e^{-n}\sqrt{2\pi n}$, we get as $n \to \infty$

$$\frac{1}{n}\log Vol(B_n(r)) = \frac{1}{n}\log\frac{\pi^n r^{2n}}{n!}$$
$$\to \frac{1}{n}\log\frac{(\pi r^2)^n}{n^n e^{-n}\sqrt{2\pi n}}$$
$$\to \log(\pi e r^2)$$

and

$$\frac{1}{l}\log q \to E[\log\det(T\boldsymbol{H}\boldsymbol{H}^\dagger)] + \log|G(T, M)|$$
$$- M(T-M)\log\pi e\sigma^2$$
$$= M(T-M)\log\mathsf{SNR} + c_{M, M}$$

which is precisely the capacity given in Theorem 9. Therefore, the channel capacity can be interpreted as packing spheres in the product space of Grassmann manifolds, as illustrated in Fig. 5.

IV. NONCOHERENT CAPACITY: GENERAL CASE

In the previous section, we discussed the multiple-antenna fading channel with same number M of transmit antennas and receive antennas, and the coherence time $T \geq 2M$. In this section, we will study other cases with general values of M, N and T.

A. The $M > N$, $T \geq 2N$ Case

For this case

$$\boldsymbol{Y} = \boldsymbol{HX} + \boldsymbol{W} = \boldsymbol{HA\Theta} + \boldsymbol{W}$$

where $\boldsymbol{Y}, \boldsymbol{W} \in \mathcal{C}^{N \times T}$. $\boldsymbol{H} \in \mathcal{C}^{N \times M}$ has i.i.d. $\mathcal{CN}(0, 1)$ entries. $\boldsymbol{X} \in \mathcal{C}^{M \times T}$. $\boldsymbol{A}$ is an $M \times M$ diagonal matrix containing the norm of the transmitted vectors. $\boldsymbol{\Theta} \in \mathcal{C}^{M \times T}$ is i.d. unitary and is independent of $\boldsymbol{A}$.

Comparing to the case with N transmit and N receive antennas, now we have more transmit antennas. If we choose only to use N antennas to transmit, the capacity derived in Theorem 9

$$C_{N, N}(\mathsf{SNR}) = N\left(1 - \frac{N}{T}\right)\log_2\mathsf{SNR} + c_{N, N} + o(1) \text{ (b/s/Hz)}$$

is asymptotically achievable. Consequently, $C_{N, N}(\mathsf{SNR})$ is a lower bound of the capacity $C_{M, N}(\mathsf{SNR})$.

In the coherent channel, by adding more transmit antennas, although the number of degrees of freedom is not increased, the capacity increases by a constant that does not depend on SNR. This increase comes from a *diversity gain*, through averaging over more fading coefficients. Somewhat surprisingly, the following theorem shows that for the noncoherent channel at high SNR, no increase whatsoever is obtained by having the extra $M - N$ transmit antennas.

Theorem 12: If $M > N$ and the coherence time $T \geq 2N$, the high SNR capacity (b/s/Hz) is given by

$$C_{M, N}(\mathsf{SNR}) = C_{N, N}(\mathsf{SNR}) + o(1)$$

where $C_{N, N}(\mathsf{SNR})$ is given in Theorem 9. This capacity can be achieved by only using N of the transmit antennas.

Proof: See Appendix D. □

The proof is technical, but the key idea is that the number of degrees of freedom for noncoherent communication actually *decreases* if one insists on spreading the power across more than N transmit antennas. Over a coherence time T, the number of spatial–temporal degrees of freedom available, even if the channel were known to the receiver, is NT, being limited by the number of *receive* antennas. Spreading the power across more than N transmit antennas cannot increase this number but only serves to increase the amount of channel uncertainty, as the dimension of the channel matrix H is now increased. Thus, the effective
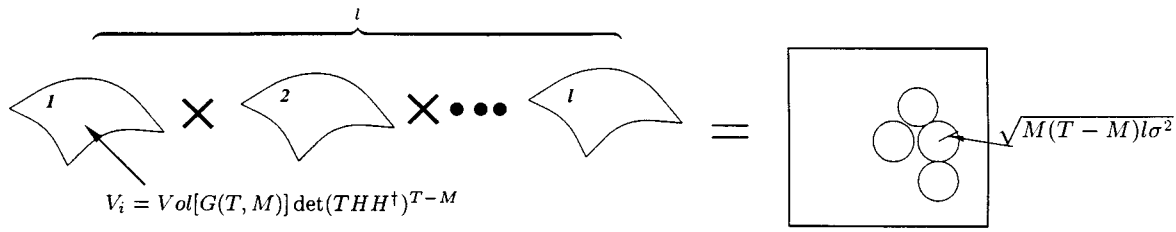
Fig. 5.   Sphere packing in noncoherent multiple-antenna channel.

degrees of freedom for noncoherent communications is actually decreased.

Let us do some heuristic calculations to substantiate this intuition. The same argument in Section III-B to make high SNR approximation of the entropy can be used

$$h(\boldsymbol{Y}) \approx h(\boldsymbol{HX}).$$

Observe that $\boldsymbol{HX}$ is i.d. We can apply Lemma 6 to yield

$h(\boldsymbol{HX})$
$$= h(\boldsymbol{C_{HX}}) + \log|G(T, N)| + (T - N)E[\log\det\boldsymbol{HA}^2\boldsymbol{H}^\dagger].$$

Condition on $\boldsymbol{X}$, $\boldsymbol{Y}$ is Gaussian with i.i.d. row vectors. The covariance of each row vector is given by $K = \boldsymbol{X}^\dagger\boldsymbol{X} + \sigma^2 I_T$. Thus, we have

$h(\boldsymbol{Y}|\boldsymbol{X})$
$$= N\sum_{i=1}^M E[\log\pi e(\|\boldsymbol{x}_i\|^2 + \sigma^2)] + N(T - M)\log\pi e\sigma^2.$$

Consider now a scheme where we use $M' \leq M$ of the transmit antennas to transmit signals with equal constant norm, and leave the rest of the antennas in silence. To keep the same total transmit power

$$\sum_{i=1}^M E[\|\boldsymbol{x}_i\|^2] = MT$$

we set $\|\boldsymbol{x}_i\|^2 = \frac{MT}{M'}$ for $i = 1, \ldots, M'$, and $\|\boldsymbol{x}_i\| = 0$ for $i = M' + 1, \ldots, M$. Let $\boldsymbol{H}_1$ contain the first M' columns of $\boldsymbol{H}$; thus

$$\boldsymbol{Y} = \frac{MT}{M'}\boldsymbol{H}_1\boldsymbol{\Theta} + \boldsymbol{W}.$$

With this input, $\boldsymbol{C_{HX}}$ has the same distribution as $\frac{MT}{M'}\boldsymbol{H}_1$; thus, the resulting mutual information is

$$I(\boldsymbol{X}; \boldsymbol{Y}) \approx h(\boldsymbol{HX}) - h(\boldsymbol{Y}|\boldsymbol{X})$$
$$= N(T - M')\log\text{SNR} + c_1(M')$$

where

$$c_1(M') = \log|G(T, N)| - N(T - M')\log\pi eM$$
$$+ (T - N)E\left[\log\det\left(\frac{MT}{M'}\boldsymbol{H}_1\boldsymbol{H}_1^\dagger\right)\right]$$
$$+ \underbrace{h\left(\frac{MT}{M'}\boldsymbol{H}_1\right) - NM'\log\pi e\left(\frac{MT}{M'} + \sigma^2\right)}_{\to 0}.$$

Observe that if $M' \geq N$, $\boldsymbol{H}_1$ has rank N with probability 1. By choosing different values of M', $c_1(M')$ only changes by a finite constant that does not depend on SNR. On the other hand, the term $N(T - M')\log\pi e\sigma^2$ yields a large difference at high SNR. The coefficient $N(T - M')$ is the number of degrees of freedom available for communication. Since NT is the total number of spatial–temporal degrees of freedom in the coherent

case, there is a *loss* of NM' degrees of freedom, increasing with M'. This loss is precisely due to the lack of knowledge of the N by M' channel matrix $\boldsymbol{H}_1$ at the receiver.

In order to maximize the mutual information at high SNR, we must choose $M' = N$ to maximize the number of degrees of freedom, which suggests the use of only N of the transmit antennas. Therefore, we conclude that if the equal constant norm input is used, the extra $M - N$ transmit antennas should be kept silent to maximize the mutual information at high SNR.

A direct generalization of the above argument results in the following statement: for a noncoherent channel with $M > N$, to maximize the mutual information at high SNR, the input should be chosen such that with probability 1, there are precisely N of the antennas transmitting a signal with strong power, i.e.,

$$\lim_{\sigma\to 0}\frac{\|\boldsymbol{x}_i\|}{\sigma} = \infty$$

and the other $M - N$ antennas have $\|\boldsymbol{x}_i\|/\sigma$ bounded. As a result, the number of degrees of freedom is not increased by having the extra $M - N$ transmit antennas.

The question now is whether the capacity can be increased by a constant amount (independent of the SNR) by allocating a *small fraction* of the transmit power to the extra antennas. Theorem 12 says no: at high SNR, one cannot do better than allocating *all* of the transmit power on only N antennas. A precise proof of this is contained in Appendix D, but some rough intuition can be obtained by going back to the coherent case. The mutual information achieved by allocating power p_i to the ith transmit antenna is given by

$$E\left[\log_2\det\left(I_N + \sum_{i=1}^M\frac{p_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger\right)\right]$$
$$= E\left[\log_2\det\left(I_N + \sum_{i=1}^N\frac{p_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger + \sum_{i=N+1}^M\frac{p_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger\right)\right]$$

where $\boldsymbol{h}_i$ is the N-dimensional vector of fading coefficients from transmit antenna i to all the N receive antennas. Since the matrix $\sum_{i=1}^N\frac{p_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger$ is full rank with probability 1, the term $\sum_{i=N+1}^M\frac{p_i}{\sigma^2}\boldsymbol{h}_i\boldsymbol{h}_i^\dagger$ will give a negligible increase in the mutual information as long as most of the power is allocated to the first N transmit antennas. The proof of Theorem 12 reveals that a similar phenomenon occurs for the noncoherent case.

One should note that the maximal degrees of freedom is obtained by using N of the M transmit antennas in both the coherent and noncoherent cases. The difference is that in the coherent case, spreading the power across all M transmit antennas retains the maximal degrees of freedom and provides a further diversity gain (reflects in a capacity increase by a constant, independent of the SNR). In contrast, there is a degrees of freedom

penalty in using more than N transmit antennas in the noncoherent case, and hence at high SNR one is forced to use only N transmit antennas even though there may be more available. Thus, no capacity gain is possible in the noncoherent case at high SNR. One should however observe that the degrees of freedom penalty is *smaller* the longer the coherence time T is, and hence the SNR level for this result to be valid is *higher* the longer T is as well. Thus, this result is meaningful at reasonable SNR levels in the regime when T is comparable to N.

B. The $M < N$, $T \geq M + N$ Case

We now consider the opposite case, when the number of receive antennas N is larger than the number of transmit antennas M. By increasing the number of the receive antennas N beyond M, intuitively, since the information-carrying object is an M-dimensional subspace, the number of degrees of freedom should still be $\dim(G(T, M)) = M(T - M)$ per coherence interval. On the other hand, the total received power is increased; hence we expect that the channel capacity to increase by a constant that does not depend on the SNR. In this section, we will argue that the equal constant norm input is optimal for $M < N$ at high SNR, and the resulting channel capacity is

$$C_{M,N}(\mathsf{SNR}) \approx M\left(1 - \frac{M}{T}\right)\log_2 \mathsf{SNR} + c_{M,N} \text{ (b/s/Hz)}$$

where

$$c_{M,N} = \frac{1}{T}\log_2 |G(T, M)| + M\left(1 - \frac{M}{T}\right)\log_2 \frac{T}{\pi e}$$
$$+ \left(1 - \frac{M}{T}\right)E[\log_2 \det \boldsymbol{H}^\dagger \boldsymbol{H}] \quad (24)$$

and

$$E[\log \det \boldsymbol{H}^\dagger \boldsymbol{H}] = \sum_{i=N-M+1}^{N} E[\log \chi_{2i}^2]$$

with χ_{2i}^2 a chi-square random variable of dimension $2i$. The number of degrees of freedom per symbol is $M(1 - M/T)$, limited by the number of *transmit* antennas.

Although the result is similar to that in Theorem 9, it turns out that some special technique has to be used for this problem.

Compared to the case where $M \geq N$ discussed in the previous sections, an important fact is that when we have less transmit antennas than receive antennas, $M < N$, we can no longer make the approximation $h(\boldsymbol{Y}) \approx h(\boldsymbol{HX})$ even at high SNR. In this case, as $\mathsf{SNR} \to \infty$, the differential entropy $h(\boldsymbol{Y})$ approaches $-\infty$ when computed in the rectangular coordinates in $\mathcal{C}^{N \times T}$. To see this, we observe that without the additive noise $\boldsymbol{W}$, the received signal $\boldsymbol{Y}_0 = \boldsymbol{HX}$ has N row vectors spanning an M-dimensional subspace. That is, the row vectors are linearly dependent of each other; therefore, $h(\boldsymbol{HX}) = -\infty$.

Similar to the coordinate change defined in (8), we can decompose $\boldsymbol{Y}_0 = \boldsymbol{HX}$ into two parts: the subspace $\Omega_{\boldsymbol{X}}$ spanned by the row vectors with dimension M and $C_{\boldsymbol{HX}} \in \mathcal{C}^{N \times M}$ to specify the position of the N row vectors inside $\Omega_{\boldsymbol{X}}$. The total number of degrees of freedom is therefore $M(T - M) + NM$.

Geometrically, we can view $\boldsymbol{HX}$ as an object on a submanifold $\mathcal{M}$ of $\mathcal{C}^{N \times T}$ with dimension $M(T - M) + NM$

$$\mathcal{M} = \{R \in \mathcal{C}^{N \times T} : \operatorname{rank}(R) = M\}.$$

Now consider the received signal $\boldsymbol{Y} = \boldsymbol{HX} + \boldsymbol{W}$, which is corrupted by the additive noise $\boldsymbol{W} \in \mathcal{C}^{N \times T}$. We can decompose $\boldsymbol{W}$ to be $\boldsymbol{W}_1$, the component on the tangent plane of $\mathcal{M}$, and $\boldsymbol{W}_2$, the component in the normal space of $\mathcal{M}$. By the argument above, we know that the dimensions of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are

$$\dim(\boldsymbol{W}_1) = M(T - M) + NM$$
$$\dim(\boldsymbol{W}_2) = NT - \dim(\boldsymbol{W}_1) = (N - M)(T - M).$$

Since $\boldsymbol{W}$ is circular symmetric, both $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ have i.i.d. $\mathcal{CN}(0, \sigma^2)$ entries.

Observe that since $\boldsymbol{HX}$ is a random object on $\mathcal{M}$, at high SNR the randomness of $\boldsymbol{Y}$ in the tangent plane of $\mathcal{M}$ is dominated by the randomness from $\boldsymbol{HX}$ rather than from the noise $\boldsymbol{W}_1$. Consequently, at high SNR, $\boldsymbol{W}_1$ has little effect on the differential entropy $h(\boldsymbol{Y})$. On the other hand, the normal space of $\mathcal{M}$ is occupied by $\boldsymbol{W}_2$ alone, which contributes a term $(N - M)(T - M)\log \pi e \sigma^2$ in $h(\boldsymbol{Y})$. Therefore, we get that as the noise level $\sigma^2 \to 0$, the differential entropy $h(\boldsymbol{Y})$ approaches $-\infty$ at the rate $(N - M)(T - M)\log \sigma^2$. In fact, by using the technique of perturbation of singular values in Appendix E, we can compute the distribution of the singular values of $\boldsymbol{Y}$, and show that at high SNR

$$h(\boldsymbol{Y}) \approx h(\boldsymbol{HAQ}) + (T - M)E[\log \det(\boldsymbol{AH}^\dagger \boldsymbol{HA})]$$
$$+ \log |G(T, M)| + (N - M)(T - M)\log \pi e \sigma^2 \quad (25)$$

where $\boldsymbol{Q} \in \mathcal{C}^{M \times M}$ is unitary i.d. matrix that is independent of $\boldsymbol{H}$ and $\boldsymbol{A}$.

To compute the conditional entropy $h(\boldsymbol{Y}|\boldsymbol{X})$, we observe that given $\boldsymbol{X}$, $\boldsymbol{Y}$ is Gaussian distributed. The row vectors are independent of each other, with the same covariance matrix $K = \boldsymbol{X}^\dagger \boldsymbol{X} + \sigma^2 I_T$. Thus, we have

$$h(\boldsymbol{Y}|\boldsymbol{X}) = N\left(\sum_{i=1}^{M} E[\log \pi e(\|\boldsymbol{x}_i\|^2 + \sigma^2)] + (T - M)\log \pi e \sigma^2\right)$$
$$\approx N\sum_{i=1}^{M} E[\log \pi e\|\boldsymbol{x}_i\|^2] + N(T - M)\log \pi e \sigma^2.$$

Combining the preceding expressions, we get

$$I(\boldsymbol{X}; \boldsymbol{Y}) = h(\boldsymbol{Y}) - h(\boldsymbol{Y}|\boldsymbol{X})$$
$$\approx \log |G(T, M)| + (T - M)E[\log \det \boldsymbol{H}^\dagger \boldsymbol{H}]$$
$$- M(T - M)\log \pi e \sigma^2 - NM\log \pi e$$
$$+ h(\boldsymbol{HAQ}) + (T - M - N)E[\log \det \boldsymbol{A}^2].$$

To maximize the mutual information, the only term that depends on the distribution of $\boldsymbol{A}$ is the last line. $\boldsymbol{HAQ}$ is an $N \times M$ matrix subject to a power constraint, thus the entropy is maximized by the matrix with i.i.d. Gaussian entries. To achieve this maximum, the input distribution has to be $P(\|\boldsymbol{x}_i\| = \sqrt{T}) = 1$ for all $i = 1, \ldots, M$. With the further assumption that $T \geq$
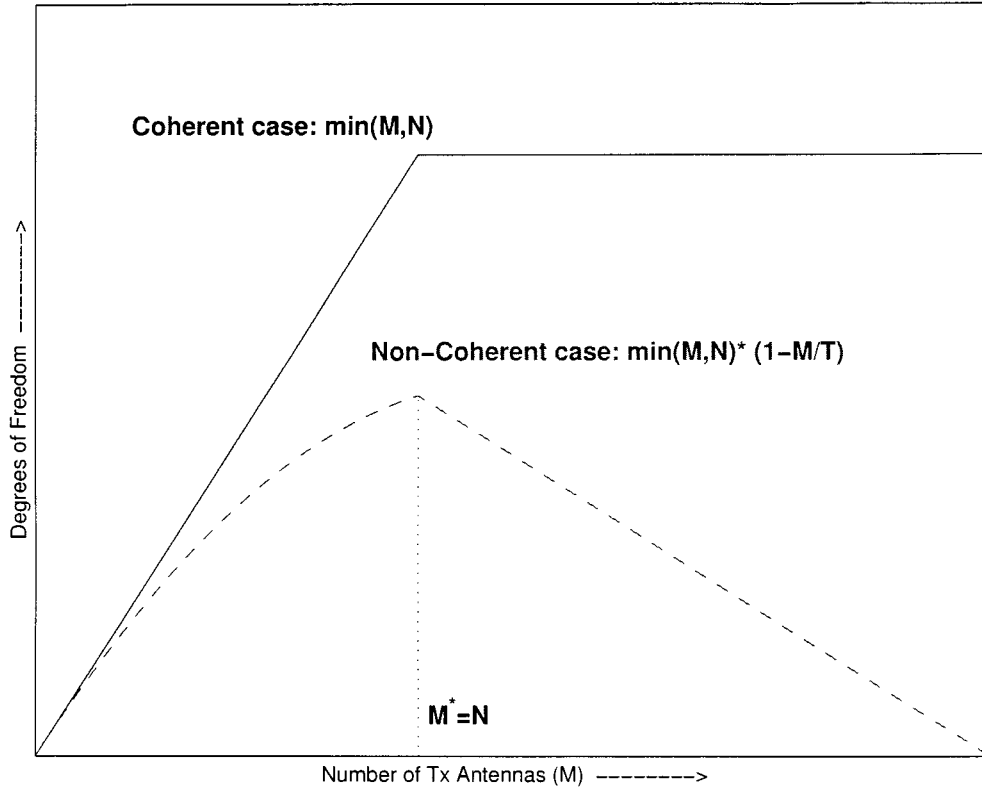
Fig. 6.   The number of degrees of freedom versus the number of Tx antennas for $T \geq 2N$.

$M + N$, $(T - M - N)E \log \det \boldsymbol{A}^2$ is also maximized by the same input distribution. Therefore, we conclude that the asymptotically optimal input distribution for the $M < N, T \geq N+M$ case is the equal constant norm input, and the maximum mutual information achieved by this input is given by

$$
\frac{1}{T} I_c(\boldsymbol{X}; \boldsymbol{Y}) \approx \log |G(T, M)| + (T - M)E[\log \det T\boldsymbol{H}^\dagger\boldsymbol{H}]
$$
$$
- M(T - M) \log \pi e \sigma^2
$$
$$
= M\left(1 + \frac{M}{T}\right)\log \mathsf{SNR} + c_{M, N} \qquad (26)
$$

where $c_{M, N}$ is defined in (24).

Comparing to $C_{M, M}(\mathsf{SNR})$ given in Theorem 9, we observe that increasing the number of receive antennas does not change the rate at which capacity increases with $\log \mathsf{SNR}$.

To make the above argument rigorous, the convergence of the approximation (25) has to be proved rigorously, which involves many technical details. As a partial proof, the following lemma shows that the approximation is an upper bound at high SNR.

Lemma 13: For the multiple-antenna channel with M transmit, N receive antennas, where $M < N$, and the coherence time $T \geq N + M$, the channel capacity (b/s/Hz) satisfies

$$
\limsup_{\mathsf{SNR}\to\infty}\left[C_{M,N}(\mathsf{SNR}) - M\left(1 - \frac{M}{T}\right)\log_2 \mathsf{SNR} - c_{M, N}\right] \leq 0
$$

where $c_{M, N}$ is defined in (24).

Proof: See Appendix E.   □

C. A Degree of Freedom View

Fig. 6 gives a bird's eye view of our results so far, focusing on the degrees of freedom attained. We fix the number of receive antennas N and the coherence time T and vary the number of transmit antennas M, and plot the (noncoherent) degrees of freedom attained by the equal constant norm input distribution on all M transmit antennas. We also assume that $T > \min\{M, N\} + N$. From the previous two subsections, the number of degrees of freedom per symbol time is

$$
\min\{M, N\}\left(1 - \frac{M}{T}\right).
$$

We also plot the number of degrees of freedom in the coherent case; this is simply given by

$$
\min\{M, N\}.
$$

It is interesting to contrast coherent and noncoherent scenarios. In the coherent channel, the number of degrees of freedom increases linearly in M and then saturates when $M \geq N$. In the noncoherent channel, the number of degrees of freedom increases sublinearly with M first, reaches the maximum at $M^* = N$, and then decreases for $M > N$. Thus, high SNR capacity for the $M > N$ case is achieved by using only N of the transmit antennas. One way to think about this is that there are two factors affecting the number of degrees of freedom in multiple-antenna noncoherent communication: the number of spatial dimensions in the system ($\min\{M, N\}$) and the amount of channel uncertainty (represented by the factor $1 - M/T$). For $M < N$, increasing M increases the

spatial dimension but introduces more channel uncertainty; however, the first factor wins out and yields an overall increase in the number of degrees of freedom For $M > N$, increasing M provides no further increase in spatial dimension but only serves to add more channel uncertainty. Thus, we do not want to use more than N transmit antennas at high SNR.

D. Short Coherence Time

In this subsection, we will study the case when $T < K + N$, where $K = \min\{M, N\}$. From the discussion in the previous sections, we know that to maximize the mutual information at high SNR, our first priority is to maximize the number of degrees of freedom. In the following, we will first focus on maximizing degrees of freedom to get some intuitive characterization of the optimal input.

First, we observe that if we have more transmit antennas than receive antennas, $M > N$, by a similar argument to that in Section IV-A we know that the mutual information per coherence interval increases with SNR no faster than $N(T - N) \log \mathsf{SNR}$. This can be achieved by using only N of the transmit antennas. In the following, we will thus only consider the system with transmit antennas no more than receive antennas, i.e., $M \leq N$. We will also assume $T > 1$.

Now suppose we use the equal constant norm input over M' of the transmit antennas, signals with power much larger than the noise.[2] Under this input, the information-carrying object is an M'-dimensional subspace ($\in G(T, M')$), thus the number of degrees of freedom available to communicate is

$$\dim(G(T, M')) = M'(T - M').$$

In Fig. 7, we plot this number as a function of M'. We observe that the the number of degrees of freedom increases with M' until $M' = \lfloor \frac{T}{2} \rfloor$, after which the number of degrees of freedom decreases. If the total number of transmit antennas $M \leq \lfloor \frac{T}{2} \rfloor$, we have to use all of the antennas to maximize the number of degrees of freedom. On the other hand, in a system with $M > \lfloor \frac{T}{2} \rfloor$, only $\lfloor T/2 \rfloor$ of the antennas should be used.

Now using the same argument as in Section IV-A, we can relax the assumption of equal constant norm input, and conclude that in a system with $M > \lfloor \frac{T}{2} \rfloor$, only $\lfloor T/2 \rfloor$ of the transmit antennas should be used to transmit signals with strong power, i.e., $\lim \|\boldsymbol{x}_i\|^2 / \sigma^2 = \infty$.

To summarize, we have that at high SNR, the optimal input must have M^* antennas transmitting signals with power much higher than the noise level, where $M^* = \min\{M, N, \lfloor \frac{T}{2} \rfloor\}$. The resulting channel capacity $C_{M, N}(\mathsf{SNR})$ satisfies

$$\underline{c} \leq C_{M, N}(\mathsf{SNR}) - M^*\left(1 - \frac{M^*}{T}\right)\log \mathsf{SNR} \leq \overline{c} \quad (27)$$

[2]Here the notion "with power much larger than the noise" means $\|\boldsymbol{x}_i\|^2 / \sigma^2 \rightarrow \infty$. For the remaining $M - M'$ antennas, signals with power comparable with the noise might be transmitted. The analysis of those weak signals, as in Appendix D, is technically hard, but it is clear that the number of degrees of freedom is not affected, since the resulting capacity gain is at most a constant independent of the SNR. Therefore, in analyzing the number of degrees of freedom we may think of the remaining $M - M'$ antennas as being silent.
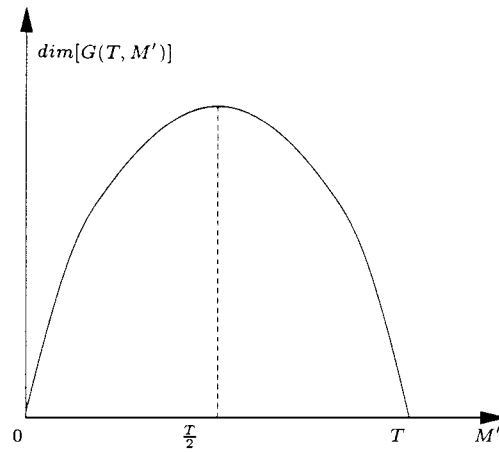

Fig. 7. Number of degrees of freedom versus number of transmit antennas.

for some constants $\overline{c}$, $\underline{c}$ that do not depend on the SNR. We observe that when the coherence time T is small, the number of useful transmit antennas is limited by T rather than the number of receive antennas N (as in Section IV-A).

Note that the result above is not as sharp as in the other cases, as the constant term is not explicitly computed. It appears that when $T < K + N$, the optimal distribution for $\boldsymbol{A}$ cannot be computed in closed form, and in general is not the equal constant norm solution.

Lemma 2 says that given the coherence time T, one needs to use at most T transmit antennas to achieve capacity. This result holds for all SNR. The above result says that at high SNR, one should in fact use no more than $\lfloor T/2 \rfloor$ transmit antennas.

V. PERFORMANCE OF A PILOT-BASED SCHEME

To communicate in a channel without perfect knowledge of the fading coefficients, a natural method is to first send a training sequence to estimate those coefficients, and then use the estimated channel to communicate. In the case when the fading coefficients are approximately time invariant (large coherence time T), one can send a long training sequence to estimate the channel accurately. However, in the case when T is limited, the choice of the length of training sequence becomes an important factor. In this section, we will study a scheme which uses T_1 symbol times at the beginning of each coherence interval to send a training sequence, and the remaining $T_2 = T - T_1$ symbol times to communicate. In the following, we will refer to the first T_1 symbol times when the pilot signals are sent as the *training phase*, and the remaining T_2 symbol times as the *communication phase*. We will describe a specific scheme, then derive the performance and compare it with the capacity results.[3]

The first key issue that needs to be addressed is: how much of the coherence interval should be allocated to channel estimation? This can be determined from a degree of freedom analysis.

[3]During the writing of this paper, we were informed by B. Hassibi of independent and related work on pilot-based schemes, in which the more general question of optimal training schemes is also addressed [8]. In this paper, we will evaluate the performance the gap between a certain pilot-based scheme and the channel capacity at high SNR.

Suppose M' of the transmit antennas is to be used in the communication phase. The total number of degrees of freedom for communication in this phase is at most

$$\min\{M', N\}(T - T_1) \tag{28}$$

the upper bound being given by the coherent capacity result (Lemma 1). On the other hand, to estimate the M' by N unknown fading coefficients, we will need at least NM' measurements at the receiver. Each symbol time yields N measurements, one at each receiver. Hence, we need a training phase of duration T_1 no smaller than M'. This represents the cost for using more transmit antennas: the more one uses, the more the time that has to be devoted to training rather than communication. Combining this with (28), the total number of degrees of freedom for communication is at most

$$\min\{M', N\}(T - M').$$

This number can be optimized with respect to M', subject to $M' \leq M$, the total number of transmit antennas. The optimal number of transmit antennas M^* to use is given by

$$M^* = \min\left\{M, N, \left\lfloor \frac{T}{2} \right\rfloor\right\}$$

with the total number of degrees of freedom given by $M^*(T - M^*)$. *This is precisely the total number of degrees of freedom promised by the capacity results.*

From this degree of freedom analysis, two insights can be obtained on the optimal number of transmit antennas to use for pilot-based schemes at high SNR.

- There is no point in using more transmit antennas than receive antennas: doing so increases the time required for training (and thereby decreases the time available for communication) but does not increase the number of degrees of freedom per symbol time for communication (being limited by the minimum of the number of transmit and receive antennas).
- Given a coherence interval of length T, there is no point in using more than $T/2$ transmit antennas. Otherwise, too much time is spent in training and not enough time for communication.

These insights mirror those we obtained in the previous noncoherent capacity analysis.

We now propose a specific pilot-based scheme which achieves the optimal number of degrees of freedom of $M^*(T - M^*)$.

- In the training phase of length $T_1 = M^*$, a simple pilot signal is used. At each symbol time, only one of the antennas is used to transmit a training symbol; the others are turned off. That is, the transmitted vector at symbol time i is $[0, \ldots, 0, x_i = D, 0, \ldots, 0]^T$. The entire pilot signal X_p is thus an $M^* \times M^*$ diagonal matrix $X_p = DI_{M^*}$.
- At the end of the training phase, all of the fading coefficients are estimated using minimum mean-square estimation (MMSE).

- In the communication phase, we communicate using the estimates of the fading coefficients $\mathbf{H}$ and the knowledge on the estimation error. We choose the input distribution $\mathbf{X}_c$ to have i.i.d. Gaussian entries, subject to the power constraint.
- We normalize the total transmitted energy in one coherence interval to be M^*T. Under this normalization, SNR $= M^*/\sigma^2$. Let $D = \sqrt{rM^*}$, where r indicates the power allocation between the training phase and the communication phase. To meet the total energy constraint, the power of the communication phase is $(M^*T - r(M^*)^2)/(T - M^*)$. If $r = 1$, the same power is used in training and communication.

In the training phase, with the pilot signals described above, the received signals can be written as

$$\mathbf{Y}_p = \sqrt{rM^*}\mathbf{H} + \mathbf{W}_p$$

where $\mathbf{Y}_p, \mathbf{H}, \mathbf{W}_p \in \mathcal{C}^{N \times M^*}$, $\mathbf{H}$ contains the NM^* unknown coefficients that are i.i.d. $\mathcal{CN}(0, 1)$ distributed, and $\mathbf{W}_p$ is the additive noise with variance σ^2.

Observe that since the entries of $\mathbf{H}$ are i.i.d. distributed, each coefficient can be estimated separately

$$\mathbf{y}_{ij} = \sqrt{rM^*}\mathbf{h}_{ij} + \mathbf{w}_{ij}, \quad \text{for } i = 1, \ldots, N; j = 1, \ldots M^*.$$

Since both $\mathbf{h}_{ij}$ and $\mathbf{w}_{ij}$ are Gaussian distributed, we can perform scalar MMSE

$$\hat{\mathbf{h}}_{ij} = \frac{\sqrt{rM^*}}{rM^* + \sigma^2}\mathbf{y}_{ij}$$

and the estimates $\hat{\mathbf{h}}_{ij}$ are independent of each other, each entry having variance

$$\hat{\sigma}_h^2 = E[|\hat{\mathbf{h}}_{ij}|^2] = \frac{rM^*}{rM^* + \sigma^2}.$$

The estimation error $\epsilon_{ij} = \mathbf{h}_{ij} - \hat{\mathbf{h}}_{ij}$ is Gaussian distributed with zero mean and the variance

$$E[|\epsilon_{ij}|^2] = \frac{\sigma^2}{rM^* + \sigma^2}.$$

Also, ϵ_{ij}'s are independent of each other.

In the communication phase, the channel can be written as

$$\begin{aligned} \mathbf{Y}_c &= \mathbf{H}\mathbf{X}_c + \mathbf{W}_c \\ &= \hat{\mathbf{H}}\mathbf{X}_c + (\mathbf{H} - \hat{\mathbf{H}})\mathbf{X}_c + \mathbf{W}_c, \end{aligned}$$

where $\mathbf{X}_c \in \mathcal{C}^{M^* \times T_2}$ has i.i.d. $\mathcal{CN}(0, \frac{T - rM^*}{T - M^*})$ entries. Define

$$\tilde{\mathbf{W}}_c = (\mathbf{H} - \hat{\mathbf{H}})\mathbf{X}_c + \mathbf{W}_c$$

as the equivalent noise in this estimated channel, one can check that the entries of $\tilde{\mathbf{W}}_c$ are uncorrelated with each other and uncorrelated to the signal $\hat{\mathbf{H}}\mathbf{X}_c$. The variance of entries of $\tilde{\mathbf{W}}_c$ is given by

$$\tilde{\sigma}^2 = \sigma^2 + M^*E[\epsilon_{ij}^2]\frac{T - rM^*}{T - M^*}.$$

The mutual information $I(\mathbf{X}_c; \mathbf{Y}_c)$ of this estimated channel is difficult to compute since the equivalent noise $\tilde{\mathbf{W}}_c$ is not
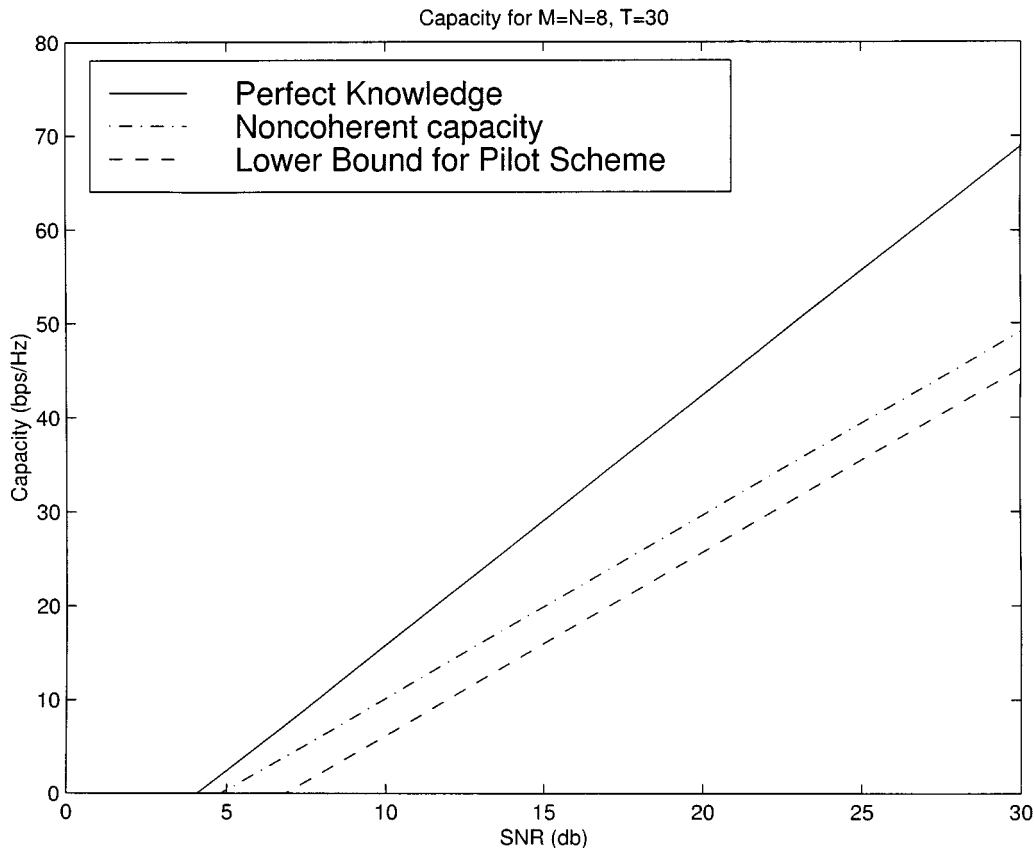
Fig. 8.   Comparison of pilot based scheme versus the noncoherent capacity.

Gaussian distributed. However, since $\tilde{\boldsymbol{W}}_c$ has uncorrelated entries and is uncorrelated to the signals $\hat{\boldsymbol{H}}\boldsymbol{X}$, it has the same first- and second-order moments as AWGN with variance $\tilde{\sigma}^2$. Therefore, if we replace $\tilde{\boldsymbol{W}}_c$ by the AWGN with the same variance, the resulting mutual information is a lower bound of $I(\boldsymbol{X}_c; \boldsymbol{Y}_c)$. Now since $\hat{\boldsymbol{H}}$ has i.i.d. Gaussian distributed entries, and is known to the receiver, this lower bound can be computed by using the result for channel with perfect knowledge of the fading coefficients, given in (4)

$$ I(\boldsymbol{X}_c; \boldsymbol{Y}_c) \geq M \log \frac{\tilde{\mathsf{SNR}}}{M^*} + \sum_{i=N-M^*+1}^{N} E[\log \chi_{2i}^2], $$

where the new SNR

$$ \tilde{\mathsf{SNR}} = \frac{M^* \frac{T-rM^*}{T-M^*} \hat{\sigma}_h^2}{\tilde{\sigma}^2} $$

$$ = \frac{\frac{r(M^*)^2(T-rM^*)}{(rM^*+\sigma^2)(T-M^*)}}{\sigma^2 + \frac{M^*\sigma^2(T-rM^*)}{(rM^*+\sigma^2)(T-M^*)}} $$

$$ = \frac{r(M^*)^2(T-rM^*)}{\sigma^2(rM^*+\sigma^2)(T-M^*) + M^*\sigma^2(T-rM^*)} $$

$$ \rightarrow \frac{r(T-rM^*)}{T+r(T-2M^*)} \frac{M^*}{\sigma^2}. $$

The last limit is taken at high SNR. We define

$$ \beta = \frac{\tilde{\mathsf{SNR}}}{\mathsf{SNR}} = \frac{r(T-rM^*)}{T+r(T-2M^*)} $$

as the SNR loss.

Thus, the lower bound of the mutual information (b/s/Hz) achieved in this pilot based scheme is given by

$$ I(\boldsymbol{X}_c; \boldsymbol{Y}_c) \geq \frac{T-M^*}{T} C_{\text{coherent}}(\beta \, \mathsf{SNR}) + o(1). \qquad (29) $$

This achieves exactly the optimal number of degrees of freedom $M^*(T - M^*)$, as claimed.

We can find the tightest bound by optimizing over the power allocation r to maximize $\tilde{\mathsf{SNR}}$. We obtain

$$ r^* = \frac{\sqrt{(M^*)^2 T^2 + M^* T^2 (T-2M^*)} - M^* T}{M^*(T-2M^*)}. $$

Now we conclude that by using the pilot based scheme described in this section, we can achieve a mutual information that increases with SNR at rate $\frac{M^*(T-M^*)}{T} \log_2 \mathsf{SNR}$ (b/s/Hz), which differs from the channel capacity only by a constant that does not depend on SNR.

The lower bound of mutual information for this scheme (29) is plotted in Fig. 8, in comparison to the noncoherent capacity derived in Theorem 9. The coherent capacity is also plotted.

Corresponding to Corollary 11, we take the large system limit by letting both M^* and T increase to ∞, but keep $\gamma = M^*/T$ fixed. Notice that the choice of r^* and the resulting SNR loss β only depend on the ratio M^*/T; thus, the resulting mutual information increases linearly with M^*, and at large M^* and high SNR

$$ \frac{1}{M^*} I(\boldsymbol{X}_c; \boldsymbol{Y}_c) \geq (1-\gamma) \log \frac{\beta}{e} \mathsf{SNR}. $$

VI. LOW SNR REGIME

In this paper, we have focused almost exclusively on the high SNR regime. It is interesting to contrast the results with the situation in the low SNR regime. First, we observe that

$$
\begin{aligned}
C_{M,N}(\mathsf{SNR}) &\le C_{\text{coherent}}(\mathsf{SNR}) \\
&= E\left[\log_2 \det\left(I_N + \frac{\mathsf{SNR}}{M} \boldsymbol{H}\boldsymbol{H}^\dagger\right)\right] \\
&\le \log_2 \det\left(I_N + \frac{\mathsf{SNR}}{M} E[\boldsymbol{H}\boldsymbol{H}^\dagger]\right) \\
&= N\log_2(1 + \mathsf{SNR}),
\end{aligned}
$$

where the second inequality follows from the concavity of the $\log\det$ function and Jensen's inequality. Hence,

$$
\limsup_{\mathsf{SNR}\to 0} \frac{C_{M,N}(\mathsf{SNR})}{\mathsf{SNR}} \le N\log_2 e.
$$

This upper bound can be asymptotically achieved by allocating all the transmit power on the first symbol of each coherence interval and on only one transmit antenna. The receiver adds up (noncoherently) the received signals from each of the receive antennas. This reduces the multiple-antenna channel with $T > 1$ to a single-antenna Rayleigh-fading channel with $T = 1$ and N times the received SNR per antenna. As is well known, the low SNR capacity of such a channel is $\mathsf{SNR}\cdot N\log_2 e$, achieving the above upper bound. (See, for instance, [9, Example 3].) Thus,

$$
\lim_{\mathsf{SNR}\to 0} \frac{C_{M,N}(\mathsf{SNR})}{\mathsf{SNR}} = N\log_2 e \qquad \text{(b/s/Hz)}.
$$

The above analysis shows that the noncoherent and coherent capacities are asymptotically equal at low SNR. Hence, in the low SNR regime, to a first order there is no capacity penalty for not knowing the channel at the receiver, unlike in the high SNR regime. Moreover, in the low SNR regime, the performance gain from having multiple antennas comes to a first order from the increase in total received power by having multiple *receive* antennas. In particular, multiple *transmit* antennas afford no performance improvement. This is in sharp contrast to the high SNR regime, where the first-order performance gain comes from the increase in degrees of freedom due to having multiple transmit *and* receive antennas. This observation is consistent with the well-known fact that a system is power-limited in the low SNR regime but degree-of-freedom-limited in the high SNR regime. Note, however, that multiple transmit antennas do yield a second-order improvement in performance at low SNR [13].

The low SNR noncoherent capacity of the multiple antenna channel is the same as that of a single-antenna Rayleigh-fading channel. As is well known, the low SNR capacity of such a channel is achieved by using a very peaky input signal, zero most of the time, and takes on a very large value with very small prbability. Thus, in the low SNR regime, information in the input $\boldsymbol{X} = \boldsymbol{A}\Theta$ to the multiple-antenna channel is in fact conveyed solely in the magnitude $\boldsymbol{A}$ and not in the subspace Ω_X

at all. This is of course just the opposite of the situation in the high SNR regime.

VII. CONCLUSION

In this paper, we studied the capacity of the noncoherent multiple-antenna channel. We used the model that assumes no prior knowledge of the channel at either the transmitter or the receiver end, but assumes that the fading coefficients remain constant for a coherence interval of length T symbol times. Under this assumption, we conclude that a system with M transmit and N receive antennas has $M^*(1 - M^*/T)$ degrees of freedom per symbol time to communicate, where $M^* = \min\{M, N, \lfloor T/2 \rfloor\}$. To utilize these degrees of freedom, the optimal strategy at high SNR and when $T \ge \min\{M, N\} + N$ is to transmit orthogonal vectors at M^* of the transmit antennas with constant equal norms, and use the subspace spanned by those vectors to carry information. The resulting channel capacity is explicitly computed as

$$
C_{M,N}(\mathsf{SNR}) = M^*\left(1 - \frac{M^*}{T}\right)\log_2 \mathsf{SNR} + c_{\min\{M,N\},N} + o(1)
$$

where $c_{K,N}$ is a constant given in (24). This expression can be interpreted as sphere packing in the Grassmann manifold. We also showed that the performance achieved by a training-based scheme is within a constant of the capacity, independent of the SNR.

We observe that having more transmit antennas than receive antennas provides no capacity gain at high SNR, while having more receive antennas does yield a capacity gain, but will not increase the number of degrees of freedom. To maximize the number of degrees of freedom in a channel with given coherence time T, the optimal number of transmit antennas is $\lfloor T/2 \rfloor$, and the number of receive antennas should be no less than $\lfloor T/2 \rfloor$.

The noncoherent communication scheme suggested by the capacity result makes no effort to estimate the channel coefficients, but uses the directions that are not affected by those coefficients to communicate. Namely, it communicates on the Grassmann manifold. However, after detecting the transmitted subspace, the receiver can always find out the directions of the transmitted vectors inside the subspace from the transmitted codeword, and perform an estimation on the fading coefficients.

APPENDIX A
COORDINATE CHANGE DEFINED BY MATRIX TRANSFORMATIONS

Differential entropies are coordinate dependent. Just as the differential entropy of a scalar random variable or a random vector can be computed in different coordinates, such as rectangular and polar coordinates, the entropy of a random matrix can be computed in different coordinates defined by standard matrix transformations. It is a widely used method in multivariate statistical analysis to view matrix transformations as coordinate changes. Research using this method can be found as early as in the 1920s. Anderson [10] provided a comprehensive overview of the field. Detailed discussions can also be found in [11], [7]. In this appendix, we will briefly summarize some of the results that are relevant to this paper.

We will start by studying the LQ decomposition of a complex matrix $R \in \mathcal{C}^{M \times N}$ for $M \leq N$

$$R = LQ \quad (30)$$

where $L \in \mathcal{C}^{M \times M}$ is a lower triangular matrix and $Q \in \mathcal{C}^{M \times N}$ is a unitary matrix, i.e., $QQ^T = I_M$. To assure that the map is one-to-one, we restrict L to have real nonnegative diagonal entries.[4]

Observe that L has $\frac{M(M-1)}{2}$ complex entries and M real entries; thus, the set of all lower triangular matrices with real nonnegative diagonals $\mathcal{L}$ has M^2 real dimensions. The number of degrees of freedom in the unitary matrix Q is $\dim(S(N, M)) = 2NM - M^2$ (real) . We observe that the total number of degrees of freedom in the right-hand side of (30) matches that of the left-hand side. In fact, the map $R \to (L, Q)$

$$\mathcal{C}^{M \times N} \to \mathcal{L} \times S(T, M)$$

is a coordinate change.

We are interested in the Jacobian of this coordinate change. This is best expressed in terms of differential forms. If we write the differentials of R, L, Q as (dR), (dL) and (dQ), respectively, then the Jacobian of this coordinate change is given by

$$\left| \frac{(dR)}{(dL)(dQ)} \right|.$$

The symbols "$(d\cdot)$" has different definitions for different kinds of matrices. For detailed discussions, please refer to [11]. From [11], we have

$$(dR) = \prod_{i=1}^{M} l_{ii}^{2(N-i)+1}(dL)(dQ). \quad (31)$$

Thus, $\left| \prod_{i=1}^{M} l_{ii}^{2(N-i)+1} \right|$ is the Jacobian of the coordinate change (30).

In the following, we will quote the Jacobian of some standard complex matrix transformations from [7], and use them to derive the Jacobian of the singular value decomposition (SVD).

Eigenvalue Decomposition

$$H = U \Lambda U^{\dagger}$$

where $H \in \mathcal{C}^{M \times M}$ is a Hermitian matrix. Λ is a diagonal matrix containing the eigenvalues. U is unitary

$$(dH) = \prod_{i<j}(\lambda_i - \lambda_j)^2(d\Lambda)(dU). \quad (32)$$

Cholesky Decomposition

$$S = LL^{\dagger}$$

where $S \in \mathcal{C}^{M \times M}$ is a Hermitian matrix, $L \in \mathcal{C}^{M \times M}$ is lower triangular with real nonnegative diagonals

$$(dS) = 2^M \prod_{i=1}^{M} l_{ii}^{2(M-i)+1}(dL). \quad (33)$$

[4]Different authors may treat the nonuniqueness of matrix factorizations in different ways, which leads to a different constant in the resulting Jacobian.

SVD of a Complex Matrix

$$R = U \Sigma V^{\dagger}$$

where $R \in \mathcal{C}^{M \times N}$, for $M \leq N$. $\Sigma \in \mathcal{C}^{M \times M}$ is a diagonal matrix containing the singular values. $U \in \mathcal{C}^{M \times M}$ and $V \in \mathcal{C}^{N \times M}$ are unitary matrices.

The Jacobian of this coordinate change is not given in [7], but can easily be derived by expressing the SVD in terms of the following composition of transformations:

$$R \overset{LQ}{\Longrightarrow} (L, Q) \overset{S=LL^{\dagger}}{\Longrightarrow} (S, Q) \overset{S=U\Lambda U^{\dagger}}{\Longrightarrow} (U, \Lambda, Q).$$

Notice that U, the eigenvectors of S, are the left eigenvectors of R, and the eigenvalues of S are the square of the singular values of R. We have

$$(dR) = \prod_{i=1}^{M} l_{ii}^{2(N-i)+1}(dL)(dQ) \quad \text{by (31)}$$

$$= \frac{1}{2^M} \prod_{i=1}^{M} l_{ii}^{2(N-M)}(dS)(dQ) \quad \text{by (33)}$$

$$= \frac{1}{2^M} \prod_{i=1}^{M} l_{ii}^{2(N-M)} \prod_{i<j}(\lambda_i - \lambda_j)^2(d\Lambda)(dU)(dQ)$$
$$\text{by (32)}$$

$$= \prod_{i<j}(\sigma_i^2 - \sigma_j^2)^2 \prod_{i=1}^{M} \sigma_i^{2(N-M)+1}(d\Sigma)(dU)(dV). \quad (34)$$

In the last step, we used $(dQ) = (dV)$ since $Q = VP$ where P is an $M \times M$ unitary matrix. In the following, we will write the Jacobian of SVD as

$$J_{N,M}(\sigma_1, \ldots, \sigma_M) = \prod_{i<j}(\sigma_i^2 - \sigma_j^2)^2 \prod_{i=1}^{M} \sigma_i^{2(N-M)+1}.$$

APPENDIX B
PROOF OF LEMMA 8

(In the sequel, we use k, k_1, etc., to denote constants that do not depend on the background noise power σ^2. Their definitions though may change in different parts of the proof.)

To prove the lemma by contradiction, we need to show that $\forall \epsilon, \delta > 0, \exists \sigma_0^2 > 0$, such that for any $\sigma^2 \leq \sigma_0^2$, any input that satisfies

$$P\left(\frac{\sigma}{\|\boldsymbol{x}_i\|} > \delta \right) > \epsilon \quad (35)$$

for some i, cannot be the optimal input. It suffices to construct another input distribution that achieves a higher mutual information, while satisfying the same power constraint.

Our proof will be outlined as follows.

1) We first show that in a system with M transmit and N receive antennas, if $M \leq N$, and the coherence time $T \geq M + N$, there exists a finite constant $k_1 < \infty$ such that for any fixed input distribution of $\boldsymbol{X}$

$$I(\boldsymbol{X}; \boldsymbol{Y}) \leq k_1 + M(T - M) \log \text{SNR}.$$

That is, the mutual information increases with SNR at a rate no higher than $M(T - M) \log \text{SNR}$.

2) Under the same assumptions, if we only choose to send a signal with strong power in M' of the transmit antennas, that is, if

$$P(\|\boldsymbol{x}_i\| \le k_2\sigma) = 1, \qquad \text{for } i = M' + 1, \ldots, M$$

and some constant k_2, we show that the mutual information increases with SNR at rate no higher than $M'(T - M') \log \text{SNR}$. This generalizes the result in the first step: even allowing $M - M'$ antennas to transmit weak power, the rate that the mutual information increases with SNR is not affected.

3) We show that if an input distribution satisfies (35), i.e., it has a positive probability that $\|\boldsymbol{x}_i\| \le k_3\sigma$, the mutual information achieved increases with SNR at rate strictly lower than $M(T - M) \log \text{SNR}$.

4) We show that for a channel with the same number M of transmit and receive antennas, by using the constant equal norm input $P(\|\boldsymbol{x}_i\| = \sqrt{T}) = 1$ for all i, the mutual information increases with SNR at rate $M(T - M) \log \text{SNR}$. Hence, any input distribution that satisfies (35) yields a mutual information that increases at a lower rate than a constant equal norm input, and thus is not optimal when σ^2 is small enough.

Step 1): For a channel with M transmit and N receive antennas, if $M < N$ and $T \ge M + N$, we write the conditional differential entropy as

$$h(\boldsymbol{Y}|\boldsymbol{X}) = h(\boldsymbol{H}) + N \sum_{i=1}^{M} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)] \\ + N(T - M) \log \pi e \sigma^2.$$

Observe that $\boldsymbol{Y}$ is circular symmetric, i.e., the eigenvectors of $\boldsymbol{Y}$ are i.d. and independent of the singular values; we compute $h(\boldsymbol{Y})$ in the SVD coordinates by (34),

$$h(\boldsymbol{Y}) = \log |S(N, N)| + \log |S(T, N)| + h(\Sigma_{\boldsymbol{Y}}) \\ + E[\log J_{T, N}(\sigma_1, \ldots, \sigma_N)]$$

where $\Sigma_{\boldsymbol{Y}} = (\sigma_1, \ldots, \sigma_N)$ are the singular values of $\boldsymbol{Y}$. We order the singular values to have $\sigma_1 \ge \cdots \ge \sigma_N$ and write

$$h(\Sigma_{\boldsymbol{Y}}) = h(\sigma_1, \ldots \sigma_M, \sigma_{M+1}, \ldots \sigma_N) \\ \le h(\sigma_1, \ldots, \sigma_M) + h(\sigma_{M+1}, \ldots, \sigma_N).$$

Consider

$$E[\log J_{T, N}(\Sigma_{\boldsymbol{Y}})] \\ = \sum_{i=1}^{N} E\left[\log \sigma_i^{2(T-N)+1}\right] + \sum_{j<j\le N} E[\log(\sigma_i^2 - \sigma_j^2)^2] \\ = \sum_{i=1}^{M} E\left[\log \sigma_i^{2(T-N)+1}\right] + \sum_{i<j\le M} E[\log(\sigma_i^2 - \sigma_j^2)^2] \\ + \sum_{i\le M, M<j\le N} \underbrace{E[\log(\sigma_i^2 - \sigma_j^2)^2]}_{\le \log \sigma_i^4}$$

$$+ \sum_{i=M+1}^{N} E\left[\log \sigma_i^{2(T-N)+1}\right] \\ + \sum_{M<i<j\le N} E[\log(\sigma_i^2 - \sigma_j^2)] \\ \le E[\log J_{N, M}(\sigma_1, \ldots, \sigma_M)] \\ + E[\log J_{T-M, N-M}(\sigma_{M+1}, \ldots, \sigma_N)] \\ + 2(T - M) \sum_{i=1}^{M} E[\log \sigma_i^2].$$

We define

$$\boldsymbol{C}_1 = \boldsymbol{U}_1 \Sigma_1 \boldsymbol{V}_1^{\dagger}, \qquad \text{where } \Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_M).$$

$\boldsymbol{U}_1 \in \mathcal{C}^{M \times M}$ and $\boldsymbol{V}_1 \in \mathcal{C}^{N \times M}$ are i.d. unitary matrices. $\boldsymbol{U}_1$, $\boldsymbol{V}_1$, Σ_1 are independent of each other. Similarly,

$$\boldsymbol{C}_2 = \boldsymbol{U}_2 \Sigma_2 \boldsymbol{V}_2^{\dagger}, \qquad \text{where } \Sigma_2 = \text{diag}(\sigma_{M+1}, \ldots, \sigma_N).$$

$\boldsymbol{U}_2 \in \mathcal{C}^{(N-M) \times (N-M)}$ and $\boldsymbol{V}_2 \in \mathcal{C}^{(T-M) \times (N-M)}$ are i.d. unitary matrices. $\boldsymbol{U}_2, \boldsymbol{V}_2$, and Σ_2 are independent of each other. Consider the differential entropy of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$

$$h(\boldsymbol{C}_1) = \log |S(M, M)| + \log |S(N, M)| \\ + h(\sigma_1, \ldots, \sigma_M) + E[\log J_{N, M}(\sigma_1, \ldots, \sigma_M)]$$
$$h(\boldsymbol{C}_2) = \log |S(N - M, N - M)| \\ + \log |S(T - M, N - M)| + h(\sigma_{M+1}, \ldots, \sigma_N) \\ + E[\log J_{T-M, N-M}(\sigma_{M+1}, \ldots, \sigma_N)].$$

Substituting in the formula of $h(\boldsymbol{Y})$, we get

$$h(\boldsymbol{Y}) \le h(\boldsymbol{C}_1) + h(\boldsymbol{C}_2) + (T - M) \sum_{i=1}^{M} E[\log \sigma_i^2] \\ + \log |S(T, N)| + \log |S(N, N)| \\ - \log |S(N, M)| - \log |S(M, M)| \\ - \log |S(N - M, N - M)| \\ - \log |S(T - M, N - M)| \\ = h(\boldsymbol{C}_1) + h(\boldsymbol{C}_2) + (T - M) \sum_{i=1}^{M} E[\log \sigma_i^2] \\ + \log |G(T, M)|. \qquad (36)$$

Remarks: To get an upper bound of $h(\boldsymbol{Y})$, we need to bound $h(\Sigma_{\boldsymbol{Y}})$. The introduction of matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ draws a connection between the singular values and matrices with lower dimensions. In the following, we will derive tight upper bound on $h(\boldsymbol{C}_1)$ and $h(\boldsymbol{C}_2)$, and hence get the bound of $h(\boldsymbol{Y})$.

Now observe that $\boldsymbol{C}_1 \in \mathcal{C}^{M \times N}$ has bounded total power

$$E\left[\sum_{i=1}^{M} \sum_{j=1}^{N} |(\boldsymbol{C}_1)_{ij}|^2\right] \\ = \sum_{i=1}^{M} E[\sigma_i^2] = E\left[\sum_{i, j} |\boldsymbol{Y}_{ij}|^2\right] \le NT(M + \sigma^2).$$

The differential entropy of C_1 is maximized if its entries are i.i.d. Gaussian distributed with variance $\frac{T(M+\sigma^2)}{M}$, thus,

$$h(C_1) \leq NM \log \pi e \frac{T(M+\sigma^2)}{M}. \tag{37}$$

Similarly, to get an upper bound of $h(C_2)$, we need to bound the total power of C_2. Since $\sigma_{M+1}, \ldots, \sigma_N$ are the $N-M$ least singular values of Y, for any unitary matrix $Q \in \mathcal{C}^{(N-M)\times N}$, we have

$$\sum_{i=1}^{N-M} \sum_{j=1}^{T-M} |(C_2)_{ij}|^2 = \sum_{i=M+1}^{N} \sigma_i^2$$
$$\leq \text{trace}(QYY^\dagger Q^\dagger).$$

Now we write $W = W_1 + W_2$, where W_1 contains the components of the row vectors in the subspace Ω_X, and W_2 contains the perpendicular components. Notice that the subspace Ω_X is independent of W, therefore, the total power in W_2 is

$$E[\text{trace}(W_2 W_2^\dagger)] = N(T-M)\sigma^2.$$

Since $HX + W_1$ has rank M, we can find a unitary matrix $Q_0 \in \mathcal{C}^{(N-M)\times M}$ such that $Q_0(HX + W_1) = 0$. Notice that Q_0 is independent of W_2, we have

$$E\left[\sum_{i=1}^{N-M} \sum_{j=1}^{T-M} |(C_2)_{ij}|^2\right] \leq E[\text{trace}(Q_0 WW^\dagger Q_0^\dagger)]$$
$$= (N-M)(T-M)\sigma^2.$$

Again, the differential entropy $h(C_2)$ is maximized if C_2 has i.i.d. Gaussian entries

$$h(C_2) \leq (N-M)(T-M) \log \pi e \sigma^2. \tag{38}$$

Substituting (37) and (38) into (36), we get

$$h(Y) \leq \log |G(T, M)| + NM \log \pi e \frac{T(M+\sigma^2)}{M}$$
$$+ (T-M) \sum_{i=1}^{M} E[\log \sigma_i^2]$$
$$+ (N-M)(T-M) \log \pi e \sigma^2. \tag{39}$$

Combining with $h(Y|X)$, we get

$$I(x; Y) \leq \underbrace{\log |G(T, M)| + NM \log \frac{T(M+\sigma^2)}{M}}_{\alpha}$$
$$+ \underbrace{(T-M-N) \sum_{i=1}^{M} E[\log \sigma_i^2]}_{\beta}$$
$$+ \underbrace{N\left(\sum_{i=1}^{M} E[\log \sigma_i^2] - \sum_{i=1}^{M} E[\log(\|x_i\|^2 + \sigma^2)]\right)}_{\gamma}$$
$$- M(T-M) \log \pi e \sigma^2. \tag{40}$$

Now the term β is upper-bounded since

$$\sum_{i=1}^{M} \sigma_i^2 = \sum_{ij} |Y_{ij}|^2$$

thus by concavity of log function

$$\sum_{i=1}^{M} E[\log \sigma_i^2] \leq M \log\left(\frac{1}{M} \sum_{i=1}^{M} E[\sigma_i^2]\right)$$
$$= M \log \frac{NT(M+\sigma^2)}{M}.$$

For the term γ, it will be shown that

$$\sum_{i=1}^{M} E[\log \sigma_i^2] - \sum_{i=1}^{M} E[\log(\|x_i\|^2 + \sigma^2)] \leq k \tag{41}$$

for some finite constant k.

Combining this with (40), we observe that the terms α, β, and γ are all upper-bounded by constants, thus, we get the desired result in Step 1).

To prove (41), we compute the expectation of the term γ by first computing the conditional expectation given X. Observe that given $X = X$, the row vectors of Y are i.i.d. Gaussian distributed with a covariance matrix $X^\dagger X + \sigma^2 I_T$. Writing $Z \in \mathcal{C}^{N\times T}$ with i.i.d. $\mathcal{CN}(0, 1)$ entries, we have

$$(YY^\dagger | X = X) \stackrel{d}{=} Z(X^\dagger X + \sigma^2 I_T)Z^\dagger$$

where $\stackrel{d}{=}$ denotes the same distribution.

Since X can be written as $X = A\Theta_1$, where $\Theta_1 \in \mathcal{C}^{M\times T}$ is a unitary matrix, let

$$\Theta = \begin{bmatrix} \Theta_1 \\ \Theta_2 \end{bmatrix}$$

be the $T \times T$ unitary matrix completed from Θ_1. Thus, we have

$$(YY^\dagger | X = X)$$
$$\stackrel{d}{=} Z\Theta^\dagger[\text{diag}(\|x_1\|^2, \ldots, \|x_M\|^2, 0) + \sigma^2 I_T]\Theta Z^\dagger.$$

Since Z has i.i.d. $\mathcal{CN}(0, 1)$ entries, $Z\Theta^\dagger$ has the same distribution as Z. If we decompose $Z \in \mathcal{C}^{N\times T}$ into block matrices $Z = [Z_1 | Z_2]$, where $Z_1 \in \mathcal{C}^{N\times M}$, $Z_2 \in \mathcal{C}^{N\times(T-M)}$, we can write

$$(YY^\dagger | X = X) \stackrel{d}{=} Z_1(A^2 + \sigma^2 I_M)Z_1^\dagger + \sigma^2 Z_2 Z_2^\dagger.$$

Now to compute

$$\sum_{i=1}^{M} E[\log \sigma_i^2], \qquad \text{for } \sigma_i, i = 1, \ldots M$$

the largest M singular values of Y, we introduce the following lemma from [12]:

Lemma 14: If C and B are both Hermitian matrices, and if their eigenvalues are both arranged in decreasing order, then

$$\sum_{i=1}^{N} (\lambda_i(C) - \lambda_i(B))^2 \leq \|C - B\|_2^2$$

where $\|M\|_2^2 \overset{d}{=} \sum M_{ij}^2$, $\lambda_i(M)$ denotes the ith eigenvalue of matrix M.

Apply this lemma to

$$C = (YY^\dagger | X = X)$$

and

$$B = Z_1(A^2 + \sigma^2 I_M)Z_1^\dagger.$$

Observe that B has only M nonzero eigenvalues, which are precisely the eigenvalues of $B' = (A^2 + \sigma^2 I_M)Z_1^\dagger Z_1 \in \mathcal{C}^{M \times M}$. Thus, for each of the M largest eigenvalues of C, we have

$$\lambda_i(C) \leq \lambda_i(B') + \sigma^2 \|Z_2 Z_2^\dagger\|_2, \qquad \text{for } i = 1, \ldots, M.$$

Observe that Z_1 has the same distribution as H, we have that for constant $k = E[\|Z_2 Z_2^\dagger\|_2]$

$$\sum_{i=1}^{M} E[\log \sigma_i^2 | X = X]$$
$$\leq \sum_{i=1}^{M} E[\log(\lambda_i((A^2 + \sigma^2 I_M)Z_1^\dagger Z_1) + \sigma^2 \|Z_2 Z_2^\dagger\|)]$$
$$\leq \sum_{i=1}^{M} E[\log(\lambda_i((A^2 + \sigma^2 I_M)Z_1^\dagger Z_1) + \sigma^2 k)]$$
$$\leq \sum_{i=1}^{M} E[\log(\lambda_i((A^2 + \sigma^2 I_M)H^\dagger H) + \sigma^2 k)]$$
$$= E[\log \det((A^2 + \sigma^2 I_M)H^\dagger H + k\sigma^2 I_M)]$$
$$= E[\log \det H^\dagger H]$$
$$\quad + E[\log \det(A^2 + \sigma^2 I_M + k\sigma^2(H^\dagger H)^{-1})]$$

where the second inequality follows from Jensen's inequality and taking expectation over Z_2. Using the lemma again on the second term, we have

$$\sum_{i=1}^{M} E[\log \sigma_i^2 | X = X]$$
$$\leq E[\log \det H^\dagger H]$$
$$\quad + E[\log \det(A^2 + \sigma^2 I_M + k\sigma^2 \|(H^\dagger H)^{-1}\|_2 I_M)]$$
$$\leq E[\log \det H^\dagger H] + E[\log \det(A^2 + k'\sigma^2 I_M)]$$

where $k' = 1 + kE[\|(H^\dagger H)^{-1}\|_2]$ is a finite constant. This again follows from Jensen's inequality.

Now we have

$$\sum_{i=1}^{M} E[\log \sigma_i^2 | X = X] - \sum_{i=1}^{M} \log(\|x_i\|^2 + \sigma^2)$$
$$\leq E[\log \det H^\dagger H] + \sum_{i=1}^{M} \log \frac{\|x_i\|^2 + k'\sigma^2}{\|x_i\|^2 + \sigma^2}$$
$$\leq E[\log \det H^\dagger H] + k'' \tag{42}$$

where k'' is another constant. Taking expectation over X, we get (41), and that completes Step 1).

Remarks: The upper bound of the mutual information so far is tight at high SNR except that k'' is not evaluated. In the later sections, we will further refine this bound by showing that $k'' \to 0$ at high SNR, and hence get a tight upper bound.

Step 2): Assume that for $M - M' > 0$ antennas, the transmitted signal has bounded SNR, that is, $P(\|x_i\|^2 < k_1 \sigma^2) = 1$ for some constant k_1. Start from a system with only M' antennas, the extra power we send on the remaining $M - M'$ antennas will get only a limited capacity gain since the SNR is bounded. Therefore, we conclude that the mutual information must be no more than $k_2 + M'(T - M') \log \mathsf{SNR}$ for some finite constant k_2 that is uniform for all SNR levels and all input distributions.

Step 3): Now we further generalize the result above to consider the input which on some of the transmit antennas, the signal transmitted has finite SNR with a positive probability, say, $P(\|x_M\|^2 < k_1 \sigma^2) = \epsilon$. Define the event

$$E = \{\|x_M\|^2 < k_1 \sigma^2\}$$

then the mutual information can be written as

$$I(X; Y) \leq \epsilon I(X; Y | E) + (1 - \epsilon)I(X; Y | E^c) + I(E; Y)$$
$$\leq \epsilon(k_2 + (M - 1)(T - M + 1) \log \mathsf{SNR})$$
$$\quad + (1 - \epsilon)(k_3 + M(T - M) \log \mathsf{SNR}) + \log 2$$

where k_1, k_2, and k_3 are finite constants. Under the assumption that $T \geq M + N$, the resulting mutual information thus increases with SNR at rate that is strictly lower than $M(T - M) \log \mathsf{SNR}$.

Step 4): Here we will show that for the channel with the same number of transmit and receive antennas, $M = N$, the constant equal norm input $P(\|x_i\| = \sqrt{T}) = 1$ for all i, we can achieve a mutual information that increase at a rate $M(T - M) \log \mathsf{SNR}$.

Lemma 15 (Achievability): For the constant equal norm input

$$\liminf_{\sigma^2 \to 0}[I(X; Y) - f(\mathsf{SNR})] \geq 0$$

where $\mathsf{SNR} = M/\sigma^2$ and
$$f(\mathsf{SNR}) = \log |G(T, M)| + (T - M)E[\log \det HH^\dagger]$$
$$\quad + M(T - M) \log \frac{T \mathsf{SNR}}{M\pi e} \tag{43}$$
where $E[\log \det HH^\dagger] = \sum_{i=1}^{M} E \log \chi_{2i}^2$.

Proof: Consider
$$h(Y) \geq h(HX)$$
$$= h(HAQ) + \log |G(T, M)|$$
$$\quad + (T - M)E[\log \det HA^2 H^\dagger]$$
$$= h(H) + MT \log T + \log |G(T, M)|$$
$$\quad + E[\log \det HH^\dagger]$$
$$h(Y | x) = h(H) + M \sum_{i=1}^{M} E[\log(\|x_i\|^2 + \sigma^2)]$$
$$\quad + M(T - M) \log \pi e \sigma^2$$
$$\leq h(H) + M^2 \log T + M^2 \frac{\sigma^2}{T}$$
$$\quad + M(T - M) \log \pi e \sigma^2.$$

So

$$I(\boldsymbol{X}; \boldsymbol{Y}) \geq \log|G(T, M)| + (T - M)E[\log \det T\boldsymbol{H}\boldsymbol{H}^\dagger]$$
$$- M(T - M)\log \pi e\sigma^2 - N^2 \frac{\sigma^2}{T}$$
$$= f(\mathsf{SNR}) + M^2 \frac{\sigma^2}{T} \to f(\mathsf{SNR}). \qquad \square$$

Combine with the results in Step 3), for any input that does not satisfy (35), since the mutual information increases at a strictly lower rate, thus, at high SNR, they are not optimal, and thus we complete the proof of Lemma 8.

APPENDIX C
PROOF OF THEOREM 9

In Appendix B, we have already shown the following results for a system with N transmit and N receive antennas.

- The mutual information achieved by any input distribution has an upper bound (40) that increases with SNR at the rate $N(T - N)\log \mathsf{SNR}$.
- By using the constant equal norm input, mutual information of $f(\mathsf{SNR})$, as defined in (43), is achievable at high SNR, see Lemma 15.
- The optimal input must satisfy $\frac{\sigma}{\|\boldsymbol{x}_i\|} \xrightarrow{P} 0$ for all $i = 1, \ldots, N$.

To show that the channel capacity is $f(\mathsf{SNR}) + o(1)$ at high SNR, since we already have a tight lower bound achieved by the constant equal norm input, it is sufficient to show that $f(\mathsf{SNR})$ is in fact an asymptotical upper bound at high SNR. Thus, we only need to use the characterization of the optimal input given in Lemma 8 to derive an upper bound that is tighter than (40).

We first observe that with the result in Lemma 8, we can get a better bound on (42)

$$E[\log \det \boldsymbol{Y}\boldsymbol{Y}^\dagger] - \sum_{i=1}^{N} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)]$$
$$\leq E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger] + \sum_{i=1}^{N} E\left[\log \frac{\|\boldsymbol{x}_i\|^2 + k_1\sigma^2}{\|\boldsymbol{x}_i\|^2 + \sigma^2}\right]$$
$$= E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger] + \sum_{i=1}^{N} E\left[\log \frac{1 + k_1\sigma^2/\|\boldsymbol{x}_i\|^2}{1 + \sigma^2/\|\boldsymbol{x}_i\|^2}\right].$$

The second term is the expectation of a bounded continuous function of $\sigma^2/\|\boldsymbol{x}_i\|^2$, thus we can apply the limit of $\sigma^2/\|\boldsymbol{x}_i\|^2 \to 0$ and get

$$\limsup_{\sigma^2 \to 0}\left[E[\log \det \boldsymbol{Y}\boldsymbol{Y}^\dagger] - \sum_{i=1}^{N} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)]\right]$$
$$\leq E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger]. \quad (44)$$

Using this result, we have
$$E[\log \det \boldsymbol{Y}\boldsymbol{Y}^\dagger]$$
$$\leq E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger] + \sum_{i=1}^{N} E[\log(\|\boldsymbol{x}_i\|^2 + k_1\sigma^2)]$$
$$\leq E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger] + N\log\left(\frac{1}{N}\sum_{i=1}^{N} E[\|\boldsymbol{x}_i\|^2] + k_1\sigma^2\right)$$
$$= E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger] + N\log(T + k_1\sigma^2). \quad (45)$$

Now substituting (44) and (45) into (40) and noticing that we are interested in the case $M = N$, we write

$$I(\boldsymbol{X}; \boldsymbol{Y}) \leq \log|G(T, N)| + N^2 \underbrace{\log \frac{T(N + \sigma^2)}{N}}_{\to \log T}$$
$$+ (T - 2N) \cdot \underbrace{E[\log \det \boldsymbol{Y}\boldsymbol{Y}^\dagger]}_{\substack{\leq E[\log \det \boldsymbol{H}\boldsymbol{H}] + N\log(T + k_1\sigma^2) \\ \to E[\log \det T\boldsymbol{H}\boldsymbol{H}^\dagger]}}$$
$$+ N \underbrace{\left(E[\log \det \boldsymbol{Y}\boldsymbol{Y}^\dagger] - \sum_{i=1}^{N} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)]\right)}_{\leq E[\log \det \boldsymbol{H}\boldsymbol{H}^\dagger]}$$
$$- N(T - N)\log \pi e\sigma^2.$$

Combining the terms, we have
$$\limsup_{\mathsf{SNR} \to \infty}[I(\boldsymbol{X}; \boldsymbol{Y}) - f(\mathsf{SNR})] \leq 0$$
which proves the theorem.

APPENDIX D
PROOF OF THEOREM 12

Let $(\boldsymbol{x}_i^{(\sigma)}, i = 1, \ldots, M)$ be the optimal input at noise level σ^2. We order the norms to have
$$\|\boldsymbol{x}_1^{(\sigma)}\| \geq \|\boldsymbol{x}_2^{(\sigma)}\| \geq \cdots \geq \|\boldsymbol{x}_M^{(\sigma)}\|.$$
Now by the argument of Appendix B, we must have
$$\frac{\sigma}{\|\boldsymbol{x}_i^{(\sigma)}\|} \xrightarrow{P} 0, \qquad \text{for } i = 1, \ldots, N \quad (46)$$
since, other wise, the mutual information achieved increases with $\log \mathsf{SNR}$ at a rate less than $N(T - N)$, which means the lower bound $C_{N,N}(\mathsf{SNR})$ is not achievable.

As before, we write
$$h(\boldsymbol{Y}) = h(\boldsymbol{U}_{\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}\boldsymbol{Q}) + \log|G(T, N)|$$
$$+ (T - N)E[\log \det \boldsymbol{Y}\boldsymbol{Y}^\dagger]$$
$$h(\boldsymbol{Y}|\boldsymbol{X}) = N\sum_{i=1}^{M} E[\log \pi e(\|\boldsymbol{x}_i\|^2 + \sigma^2)]$$
$$+ N(T - M)\log \pi e\sigma^2.$$

Now for any input distribution $P(\boldsymbol{A})$, let $\boldsymbol{A}_1$ as the $N \times N$ diagonal matrix contain the N largest norms $\|\boldsymbol{x}_1\|, \ldots, \|\boldsymbol{x}_N\|$, and $\boldsymbol{A}_2$ is an $(M - N) \times (M - N)$ diagonal matrix with entries $\|\boldsymbol{x}_{N+1}\|, \ldots, \|\boldsymbol{x}_M\|$. Correspondingly, the partitions $\boldsymbol{H}$ and $\boldsymbol{\Theta}$, we can write
$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{\Theta} + \boldsymbol{W}$$
$$= \boldsymbol{H}_1\boldsymbol{A}_1\boldsymbol{\Theta}_1 + \boldsymbol{H}_2\boldsymbol{A}_2\boldsymbol{\Theta}_2 + \boldsymbol{W}.$$

Define $\boldsymbol{Y}_1 = \boldsymbol{H}_1\boldsymbol{A}_1\boldsymbol{\Theta}_1 + \boldsymbol{W}$. We construct input distribution P_0 from P by setting $P_0(\boldsymbol{A}_1) = P(\boldsymbol{A}_1)$ and $\boldsymbol{A}_2 = 0$. That is, we keep the distribution of the N largest norms, but set the other $M - N$ norms to 0. We observe that P_0 uses less power than P. Now we define input distribution Q such that it has the same total average power as P_0 but uses only N antennas to transmit equal constant power. To show that by using extra power on the extra $M - N$ transmit antennas, it provides no capacity gain at high SNR, we only need to compare the mutual information generated by P and Q and show that
$$\limsup_{\sigma^2 \to 0}[I_P - I_Q] \leq 0.$$

Using the expression of differential entropies above, we write

$$I_P - I_Q = h_P(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) - h_Q(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) + \Delta$$

where

$$
\begin{aligned}
\Delta = {} & (T-N)E_P[\log\det \boldsymbol{YY}^\dagger] \\
& - N\sum_{i=1}^{M} E_P[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)] \\
& - \Bigg[(T-N)E_Q[\log\det \boldsymbol{YY}^\dagger] \\
& \quad - N\sum_{i=1}^{N} E_Q[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)] \\
& \quad - N(M-N)\log\sigma^2\Bigg].
\end{aligned}
$$

From Appendix C, we know that in an $N \times N$ system, given $T \geq 2N$, the term

$$(T-N)E[\log\det \boldsymbol{YY}^\dagger] - N\sum_{i=1}^{N} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)]$$

is maximized at high SNR by a constant equal norm input. That is, if we replace the last line of the expression above by take expectation over P_0, we will get an upper bound

$$
\begin{aligned}
\Delta \leq {} & (T-N)E_P[\log\det \boldsymbol{YY}^\dagger] \\
& - N\sum_{i=1}^{M} E_P[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)] \\
& - \Bigg[(T-N)E_{P_0}[\log\det \boldsymbol{YY}^\dagger] \\
& \quad - N\sum_{i=1}^{N} E_{P_0}[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)] \\
& \quad - N(M-N)\log\sigma^2\Bigg] \\
= {} & (T-N)(E[\log\det \boldsymbol{YY}^\dagger] - E[\log\det \boldsymbol{Y}_1\boldsymbol{Y}_1]) \\
& - N\sum_{i=N+1}^{M} E\left[\log\left(1 + \frac{\|\boldsymbol{x}_i\|^2}{\sigma^2}\right)\right]
\end{aligned}
$$

where all the expectations above are taken with respect to the distribution P_0 (as will also be the case for all the random variables in the remainder of this appendix).

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ be the N eigenvalues of $\boldsymbol{Y}_1\boldsymbol{Y}_1^\dagger$. Now since $\boldsymbol{YY}^\dagger = \boldsymbol{Y}_1\boldsymbol{Y}_1^\dagger + \boldsymbol{H}_2\boldsymbol{A}_2^2\boldsymbol{H}_2^\dagger$, by Lemma 14, we know that each eigenvalue of $\boldsymbol{YY}^\dagger$ is perturbed from the corresponding λ_i by no more than $\|\boldsymbol{H}_2\boldsymbol{A}_2^2\boldsymbol{H}_2^\dagger\|_2$. Since $\|\boldsymbol{x}_{N+1}\|$ is the largest element of $\boldsymbol{A}_2$, we have that for some finite constant k_1

$$E[\log\det \boldsymbol{YY}^\dagger] \leq \sum_{i=1}^{N} E[\log(\lambda_i + k_1\|\boldsymbol{x}_{N+1}\|^2)].$$

Therefore,

$$
\begin{aligned}
\Delta \leq {} & (T-N)\sum_{i=1}^{N} E\left[\log\left(1 + k_1\frac{\|\boldsymbol{x}_{N+1}\|^2}{\lambda_i}\right)\right] \\
& - N\sum_{i=1}^{N} E\left[\log\left(1 + \frac{\|\boldsymbol{x}_{N+1}\|^2}{\sigma^2}\right)\right] \\
\leq {} & (T-N)NE\left[\log\left(1 + k_1\frac{\|\boldsymbol{x}_{N+1}\|^2}{\lambda_N}\right)\right] \\
& - NE\left[\log\left(1 + \frac{\|\boldsymbol{x}_{N+1}\|^2}{\sigma^2}\right)\right] \\
\leq {} & (T-N)NE\left[\log\left(1 + k_2\frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_N\|^2}\right)\right] \\
& - NE\left[\log\left(1 + \frac{\|\boldsymbol{x}_{N+1}\|^2}{\sigma^2}\right)\right] \\
\leq {} & (T-N)NE\left[k_2\frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_N\|^2}\right] \\
& - NE\left[\log\left(1 + \frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_N\|^2}\frac{\|\boldsymbol{x}_N\|^2}{\sigma^2}\right)\right]
\end{aligned}
$$

for a finite constant k_2.

Now we define the event $\mathcal{E} = \{\|\boldsymbol{x}_N\|^2 \geq L\sigma^2\}$. Since $\|\boldsymbol{x}_N\|$ satisfies (46), we have that for any L, $P(\mathcal{E}^c) \to 0$. It is easy to check that given $\mathcal{E}^c$, the conditional expectation $\Delta_{\mathcal{E}^c} < \infty$, thus $P(\mathcal{E}^c)\Delta_{\mathcal{E}^c}$ is arbitrarily small at high SNR, and it is sufficient to only consider Δ given $\mathcal{E}$

$$
\begin{aligned}
\Delta_{\mathcal{E}} \leq {} & (T-N)Nk_2 E\left[\frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_N\|^2}\right] \\
& - NE\left[\log\left(1 + L\frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_N\|^2}\right)\right].
\end{aligned}
$$

Consider the function

$$g(t) = (T-N)k_2 t - \log(1 + Lt).$$

It is easy to check that $g(0) = 0$. Also $g'(t) < 0$ for $t < t_0 = \frac{L-(T-N)k_2}{(T-N)k_2 L}$, and $g'(t) > 0$ for $t > t_0$. For large enough L, we have $g(1) = (T-N)k_2 - \log(1+L) < 0$, which implies that $\forall t \leq 1$, $g(t) < 0$. Using this result for $t = \|\boldsymbol{x}_{N+1}\|^2/\|\boldsymbol{x}_N\|^2 \leq 1$, we have that $\Delta_{\mathcal{E}} \leq 0$, and hence

$$\limsup_{\sigma^2 \to 0} \Delta \leq 0.$$

Furthermore, we observe for any strictly positive t, there exists a large enough L such that $g(t)$ is arbitrarily negative. This implies that if

$$P\left(\frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_N\|^2} > \epsilon\right) = \delta$$

for any $\epsilon > 0$ and $\delta > 0$, then

$$\limsup_{\sigma^2 \to 0} \Delta = -\infty.$$

Thus, we conclude that if $\limsup \Delta > -\infty$, we must have

$$\frac{\|\boldsymbol{x}_{N+1}\|^2}{\|\boldsymbol{x}_i\|^2} \xrightarrow{P} 0. \tag{47}$$

Now the matrix $\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q} \in \mathcal{C}^{N\times N}$ has limited total power. The differential entropy is maximized by a matrix with i.i.d. Gaussian entries

$$h_P(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) \leq N^2 \log \pi e \left(NT\sigma^2 + \sum_{i=1}^{M} E_P[\|\boldsymbol{x}_i\|^2] \right).$$

On the other hand, since $\boldsymbol{Q}$ is constant, an equal norm input with total transmit power $\sum_{i=1}^{N} E_P[\|\boldsymbol{x}_i\|^2]$

$$h_Q(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) \to N^2 \log \pi e \left(NT\sigma^2 + \sum_{i=1}^{N} E_P[\|\boldsymbol{x}_i\|^2] \right)$$

as $\sigma \to 0$. Thus,

$$h_P(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) - h_Q(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q})$$
$$\leq N^2 \log \left(1 + \frac{\sum_{i=N+1}^{M} E[\|\boldsymbol{x}_i\|^2]}{\sum_{i=1}^{N} E[\|\boldsymbol{x}_i\|^2]} \right) \leq k_3 < \infty \quad (48)$$

hence

$$I_P - I_Q = h_P(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) - h_Q(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) + \Delta$$
$$\leq k_3 + \Delta.$$

In order to have $k_3 + \Delta \geq 0$, use (47), and we have

$$\|\boldsymbol{x}_{N+1}\|^2 / \|\boldsymbol{x}_N\|^2 \to 0$$

in probability. Applying this result in (48), we have

$$\limsup_{\sigma^2 \to 0} h_P(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) - h_Q(\boldsymbol{U_Y}\boldsymbol{\Sigma_Y}\boldsymbol{Q}) \leq 0.$$

Combining the results we have

$$\limsup_{\sigma^2 \to 0} [I_P - I_Q] \leq 0$$

which completes the proof.

APPENDIX E
PROOF OF LEMMA 13

In Appendix B, we have shown that for a system with M transmit and N receive antennas, where $M \leq N$, if $T \geq M + N$, for any input distribution of $\boldsymbol{X}$ that satisfies (35), the mutual information achieved increases with the SNR at a rate strictly lower than $M(T - M) \log \text{SNR}$. On the other hand, by using a constant equal norm input, the mutual information is lower-bounded by $C_{M,M}(\text{SNR})$, which increases with SNR at a rate $M(T - M) \log \text{SNR}$. Therefore, we conclude that the optimal input distribution $(\boldsymbol{x}_i^{(\sigma)}, i = 1, \ldots, M)$ must satisfy

$$\frac{\sigma}{\|\boldsymbol{x}_i^{(\sigma)}\|} \xrightarrow{P} 0.$$

Similarly to the proof of Theorem 9, in the following, we will use this convergence to find a tight upper bound for the channel capacity. For simplicity, we rewrite the channel as follows:

$$\boldsymbol{Y} = \boldsymbol{HX} + \boldsymbol{W} \quad (49)$$

where $\boldsymbol{Y}, \boldsymbol{W} \in \mathcal{C}^{N\times T}$, $\boldsymbol{X} = \boldsymbol{A\Theta} \in \mathcal{C}^{M\times T}$. $\boldsymbol{H}$ is an $N \times M$ matrix with i.i.d. $\mathcal{CN}(0, 1)$ entries. We decompose into $\boldsymbol{W} = \boldsymbol{W}_1 + \boldsymbol{W}_2$, where $\boldsymbol{W}_1$ is the component of each row vectors of $\boldsymbol{W}$ in $\Omega_{\boldsymbol{X}}$, and $\boldsymbol{W}_2$ is the perpendicular component.

Now as an improvement of (38), we observe that

$$\boldsymbol{YY}^\dagger = (\boldsymbol{HX} + \boldsymbol{W}_1)(\boldsymbol{HX} + \boldsymbol{W}_1)^\dagger + \boldsymbol{W}_2\boldsymbol{W}_2^\dagger.$$

Since $\boldsymbol{HX} + \boldsymbol{W}_1$ has only rank M, we can find a unitary matrix $\boldsymbol{Q}_0 \in \mathcal{C}^{(N-M)\times N}$ such that $\boldsymbol{Q}_0(\boldsymbol{HX} + \boldsymbol{W}_1) = 0$. Therefore, we have

$$E\left[\sum_{i=1}^{N-M} \sum_{j=1}^{T-M} |(\boldsymbol{C}_2)_{ij}|^2 \right] \leq E[\text{trace}(\boldsymbol{Q}_0\boldsymbol{W}_2\boldsymbol{W}_2^\dagger\boldsymbol{Q}_0^\dagger)]$$
$$= (N - M)(T - M)\sigma^2$$

and $h(\boldsymbol{C}_2) \leq (N - M)(T - M) \log \pi e\sigma^2$.

Equation (40) thus becomes

$$I(\boldsymbol{X}; \boldsymbol{Y}) \leq \log |G(T, M)| + NM \log \frac{T(M + \sigma^2)}{M}$$
$$- M(T - M) \log \pi e\sigma^2$$
$$+ (T - M - N) \sum_{i=1}^{M} E[\log \sigma_i^2]$$
$$+ N \left(\sum_{i=1}^{M} E[\log \sigma_i^2] - \sum_{i=1}^{M} E[\log(\|\boldsymbol{x}_i\| + \sigma^2)] \right).$$

The second improvement is that from (42) we have

$$\sum_{i=1}^{M} E[\log \sigma_i^2] - \sum_{i=1}^{M} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)]$$
$$\leq E[\log \boldsymbol{H}^\dagger\boldsymbol{H}] + \sum_{i=1}^{M} E\left[\log \frac{\|\boldsymbol{x}_i\|^2 + k'\sigma^2}{\|\boldsymbol{x}_i\|^2 + \sigma^2} \right].$$

The second term is the expectation of a bounded continuous function of $\sigma^2/\|\boldsymbol{x}_i\|^2$, which converges to 0 in probability. Applying that limit we have

$$\limsup_{\sigma^2 \to 0} \sum_{i=1}^{M} E[\log \sigma_i^2] - \sum_{i=1}^{M} E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)]$$
$$\leq E[\log \det \boldsymbol{H}^\dagger\boldsymbol{H}].$$

Also, since

$$E[\log(\|\boldsymbol{x}_i\|^2 + \sigma^2)] \leq \log(E[\|\boldsymbol{x}_i\|^2] + \sigma^2)$$
$$= \log(T + \sigma^2) \to \log T$$

we have

$$\limsup_{\sigma^2 \to 0} \sum_{i=1}^{M} E[\log \sigma_i^2] \leq E[\log \det \boldsymbol{H}^\dagger\boldsymbol{H}] + \log T.$$

Combining all the results so far we have

$$\limsup_{\sigma^2 \to 0} I(\boldsymbol{X}; \boldsymbol{Y}) \leq \log |G(T, M)| - M(T - M) \log \pi e\sigma^2$$
$$+ (T - M)E[\log \det T\boldsymbol{H}^\dagger\boldsymbol{H}].$$

Substituting $\text{SNR} = M/\sigma^2$, we get the desired result.

APPENDIX F

HEURISTIC DERIVATION OF (25)

First, using the change of coordinates of SVD defined in (34), let $\sigma_1, \ldots, \sigma_N$ be the singular values of $\boldsymbol{Y}$, we write

$$
\begin{aligned}
h(\boldsymbol{Y}) = {} & h(\boldsymbol{U_Y}) + h(\boldsymbol{V_Y}) + h(\sigma_1, \ldots, \sigma_N) \\
& + E[\log |J_{T,N}(\sigma_1, \ldots, \sigma_N)|] \\
= {} & \log |S(N, N)| + \log |S(T, N)| + h(\sigma_1, \ldots, \sigma_N) \\
& + E[\log |J_{T,N}(\sigma_1, \ldots, \sigma_N)|].
\end{aligned}
$$

Now we need to compute the distribution of the singular values of $\boldsymbol{Y}$, to do that, we introduce the following lemma.

Lemma 16: For the $\boldsymbol{Y}$ given in (49), fix an input norms distribution $P(\|\boldsymbol{x}_i\|, i = 1, \ldots M)$ satisfying Lemma 8. If we order the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$, then the vector

$$
\left(\sigma_1, \sigma_2, \ldots, \sigma_M, \frac{\sigma_{M+1}}{\sigma}, \ldots, \frac{\sigma_N}{\sigma}\right)
$$
$$
\xrightarrow{d} (\mu_1, \mu_2, \ldots, \mu_N) \quad (50)
$$

as background noise level $\sigma^2 \to 0$, where $\mu_1, \ldots, \mu_M$ are the singular values of $\boldsymbol{HA} \in \mathcal{C}^{N \times M}$, and $\mu_{M+1}, \ldots, \mu_N$ is the singular value of an independent $(N - M) \times (T - M)$ matrix with i.i.d. $\mathcal{CN}(0, 1)$ entries.

This lemma can be rigorously proved. Although the proof we have is too complicated even to be included in this appendix, the intuition behind it can be briefly illustrated here. Consider the following equation with the roots $(\sigma_1^2, \ldots, \sigma_N^2)$:

$$
f(\lambda) = \det(\lambda I_N - \boldsymbol{YY}^\dagger) = 0.
$$

By the circular symmetry of the noise matrix $\boldsymbol{W}$, the random matrix $\boldsymbol{Y}$ has the same distribution as

$$
\boldsymbol{R} = \begin{bmatrix} \boldsymbol{D} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix}.
$$

Write

$$
f(\lambda) = \det(\lambda I_N - \boldsymbol{RR}^\dagger).
$$

At high SNR, we can simplify this formula by ignoring the terms with higher order of σ^2

$$
f(\lambda) \approx \det\left(\begin{bmatrix} \lambda I_M - \boldsymbol{D}^2 & -\boldsymbol{DW}_{21}^\dagger \\ -\boldsymbol{W}_{21}^\dagger \boldsymbol{D} & \boldsymbol{B} \end{bmatrix}\right)
$$

where

$$
\boldsymbol{B} = \lambda I_{N-M} - \boldsymbol{W}_{21}\boldsymbol{W}_{21}^\dagger - \boldsymbol{W}_{22}\boldsymbol{W}_{22}^\dagger.
$$

Now using Schur's identity for a determinant of block matrix

$$
\det\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \det(A)\det(D - CA^{-1}B)
$$

we get

$$
\begin{aligned}
f(\lambda) \approx {} & \det(\lambda I_M - \boldsymbol{D}^\dagger)\det(\lambda I_{N-M} - \boldsymbol{W}_{22}\boldsymbol{W}_{22}^\dagger \\
& - \boldsymbol{W}_{21}\boldsymbol{W}_{21}^\dagger + \boldsymbol{W}_{21}D(\boldsymbol{D}^\dagger - \lambda I_M)^{-1}\boldsymbol{DW}_{21}^\dagger).
\end{aligned}
$$

To find the roots of the equation $f(\lambda) = 0$, we observe that the first M roots are the entries in $\boldsymbol{D}$. Furthermore, since the other $N - M$ roots are of the order σ^2, thus they are much smaller than the entries of $\boldsymbol{D}$ at high SNR. We approximate

$(\boldsymbol{D}^2 - \lambda I_M)^{-1}$ as $\boldsymbol{D}^{-2}$, the second determinant becomes $\det(\lambda I_{N-M} - \boldsymbol{W}_{22}\boldsymbol{W}_{22}^\dagger)$. Therefore, the remaining $N - M$ eigenvalues of $\boldsymbol{YY}^\dagger$ are approximately the eigenvalues of $\boldsymbol{W}_{22}\boldsymbol{W}_{22}^\dagger$.

Lemma 16 states that the large singular values and the scaled small singular values of $\boldsymbol{Y}$ are asymptotically independent at high SNR. This justifies the following approximation for $h(\boldsymbol{Y})$:

$$
\begin{aligned}
h(\boldsymbol{Y}) = {} & \log |S(N, N)| + \log |S(T, N)| + h(\sigma_1, \ldots \sigma_N) \\
& + E[\log |J_{T,N}(\sigma_1, \ldots, \sigma_N)|] \\
= {} & \log |S(N, N)| + \log |S(T, N)| + h(\sigma_1, \ldots \sigma_M) \\
& + h(\sigma_{N+1}, \ldots, \sigma_M) \\
& + E[\log |J_{T,N}(\sigma_1, \ldots, \sigma_N)|]. \quad (51)
\end{aligned}
$$

Now letting $\boldsymbol{Q}$ be an i.d unitary $M \times M$ matrix independent of $\boldsymbol{HA}$, consider

$$
\begin{aligned}
h(\boldsymbol{HAQ}) = {} & \log |S(N, N)| + \log |S(N, M)| \\
& + h(\sigma_1, \ldots, \sigma_M) \\
& + E[\log |J_{N,M}(\sigma_1, \ldots, \sigma_M)|]
\end{aligned}
$$

where, by Lemma 16, $\sigma_1, \ldots, \sigma_M$ are identical as in (51).

Also, we write

$$
\begin{aligned}
h(\boldsymbol{W}_{22}) = {} & (N - M)(T - M)\log \pi e \sigma^2 \\
= {} & \log |S(N - M, N - M)| \\
& + \log |S(T - M, N - M)| \\
& + h(\sigma_{M+1}, \ldots \sigma_N) \\
& + E[\log |J_{T-M, N-M}(\sigma_{M+1}, \ldots, \sigma_N)|].
\end{aligned}
$$

Again, by Lemma 16, the singular values of $\boldsymbol{W}_{22}$ have approximately the same distribution as the $N - M$ smallest singular values of $\boldsymbol{Y}$ at high SNR, thus they are denoted as $\sigma_{M+1}, \ldots, \sigma_N$ in (51). Combining the three equations, we get

$$
\begin{aligned}
h(\boldsymbol{Y}) \approx {} & h(\boldsymbol{HAQ}) + h(\boldsymbol{W}_{22}) \\
& \left.\begin{aligned} & + E[\log |J_{T,N}(\sigma_1, \ldots, \sigma_N)|] \\ & - E[\log |J_{N,M}(\sigma_1, \ldots, \sigma_M)|] \\ & - E[\log \det |J_{T-M, N-M}(\sigma_{M+1}, \ldots \sigma_N)|] \end{aligned}\right\} \alpha \\
& \left.\begin{aligned} & + \log |S(N, N)| + \log |S(T, N)| \\ & - \log |S(N, M)| - \log |S(M, M)| \\ & - \log |S(N - M, N - M)| \\ & - \log |S(T - M, N - M)| \end{aligned}\right\} \beta.
\end{aligned}
$$

Substituting the definition of $J_{T,N}$ in (34), we get

$$
\begin{aligned}
\text{Term } \alpha = {} & 2 \sum_{i<j\leq N} E[\log(\sigma_i^2 - \sigma_j^2)] \\
& + (2(T - N) + 1) \sum_{i=1}^{N} E[\log \sigma_i] \\
& - 2 \sum_{i<j\leq M} E[\log(\sigma_i^2 - \sigma_j^2)] \\
& - (2(N - M) + 1) \sum_{i=1}^{M} E[\log \sigma_i] \\
& - \sum_{M<i<j\leq N} E[\log(\sigma_i^2 - \sigma_j^2)] \\
& - (2(T - N) + 1) \sum_{i=M+1}^{N} E[\log \sigma_i^2].
\end{aligned}
$$

Since the first M singular values are much larger than the last $M - N$ values, we have

$$
\begin{aligned}
&2 \sum_{i<j\leq N} \log(\sigma_i^2 - \sigma_j) - 2 \sum_{j<i\leq M} \log(\sigma_i^2 - \sigma_j^2) \\
&\quad - 2 \sum_{M<i<j\leq N} \log(\sigma_i^2 - \sigma_j^2) \\
&= 2 \sum_{i\leq M,\, j>M} \log(\sigma_i^2 - \sigma_j^2) \\
&\approx 2(N-M) \sum_{i=1}^{M} \log \sigma_i^2.
\end{aligned}
$$

Thus, the term α becomes

$$
\begin{aligned}
\text{Term } \alpha &\approx 2(N-M) \sum_{i=1}^{M} \log \sigma_i^2 \\
&= (T-M)E[\log \det(\boldsymbol{AH^{\dagger}HA})].
\end{aligned}
$$

Also, substituting into the definition of $|S(T, N)|$ in (6), we have

$$
\text{Term } \beta = \log |G(T, M)|.
$$

REFERENCES

[1] G. J. Foschini, "Layered space–time architecture for wireless communication in a fading environment when using multi-element antennas," *AT&T Bell Labs. Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[2] G. Foschini and M. Gans, "On limits of wireless communications in fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, pp. 311–335, 1998.

[3] I. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommun.*, vol. 10, pp. 585–595, Nov./Dec. 1999.

[4] T. Marzetta and B. Hochwald, "Capacity of mobile multiple-antenna communication link in a Rayleigh flat-fading environment," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999.

[5] B. Hochwald and T. Marzetta, "Unitary space–time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 46, pp. 543–565, Mar. 2000.

[6] W. M. Boothby, *An Introduction to Differential Manifolds and Riemannian Geometry*, 2nd ed. San Diego, CA: Academic, 1986.

[7] A. Edelman, "Eigenvalues and condition numbers of random matrices," Ph.D. dissertation, MIT, Cambridge, MA, 1989.

[8] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links," *IEEE. Trans. Inform. Theory*, submitted for publication.

[9] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1019–1030, Sept. 1990.

[10] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.

[11] R. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982.

[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[13] S. Verdú, "Spectral efficiency in the wide-band regime," *IEEE Trans. Inform. Theory*, submitted for publication.