Fundamental Limits of Training-Based Multiuser MIMO Systems

Xiaojun Yuan, Senior Member, IEEE, Congmin Fan, Student Member, IEEE, and Ying Jun (Angela) Zhang, Senior Member, IEEE

Abstract—In this paper, we endeavour to seek a fundamental understanding of the potentials and limitations of training-based multiuser multiple-input multiple-output (MIMO) systems. In a multiuser MIMO system, users are geographically separated. So, the near-far effect plays an indispensable role in channel fading. The existing optimal training design for conventional MIMO does not take the near-far effect into account, and thus is not applicable to a multiuser MIMO system. In this work, we use the majorization theory as a basic tool to study the tradeoff between the channel estimation quality and the information throughput. We establish tight upper and lower bounds of the throughput, and prove that the derived lower bound is asymptotically optimal.

Index Terms—Training-based multiuser MIMO, throughput maximization, massive MIMO

I. INTRODUCTION

MULTIPLE antenna (a.k.a. MIMO: multiple-input multiple-output) techniques have been extensively studied to improve the spectral efficiency of mobile communication systems, and are envisioned to be ubiquitously supported to accommodate the exponential growth of future wireless service demands. MIMO communications, however, requires the knowledge of the channel state information (CSI) at the transmitter for precoding and at the receiver for signal detection. A common approach is to allocate a certain amount of time and power resources to construct and transmit training signals for acquiring CSI. The impact of CSI acquisition on the overall performance of a MIMO system has been investigated under various performance measures, such as channel minimum mean-square error (MMSE), bit error rate (BER), and channel input output mutual information (MI) [1]–[7]. Among these measures, MI is advantageous in that it characterizes the fundamental tradeoff between achieving high-quality channel estimate and the information throughput. One on hand, to achieve a high-quality channel estimate, more time and power resources should be allocated for training, leaving the data transmission with little time and power. On the other hand, if too little resources are allocated to training, information throughput deteriorates due to channel mismatch. The exact throughput characterization of the training-based MIMO system is difficult, but a tight MI lower bound was derived in [1], [2]. The authors in [1] also discussed the tradeoff from the perspective of degrees of freedom (DoF) [1]. Later, Coldrey et al. established the optimal tradeoff by assuming independent and identically distributed (i.i.d.) Rayleigh fading and exploiting the rotational invariance property of an i.i.d. Gaussian channel matrix [5].

Recently, the incorporation of MIMO into multiuser cellular systems has attracted much research interest, especially in a massive MIMO setup where users communicate with a base station (BS) equipped with a large-scale antenna array [8]–[11]. A multiuser MIMO system can be treated as a virtual MIMO system without cooperation at the transmitter side. It exhibits some new features compared with conventional MIMO. First, geographically separated users in general experience significantly different large-scale fading caused by path loss and shadowing. That is, the near-far effect is indispensable in modelling a multiuser MIMO system. Second, the total transmission power of a MIMO system is usually constrained, while that of a multiuser MIMO system scales with the number of active users. Third, large antenna arrays may be deployed at BSs, focusing energy into ever-smaller spatial regions to bring huge improvements in system throughput and energy efficiency; but at the same time, large antenna arrays creates a lot more channel links than ever before, which imposes a heavy burden on the acquisition of CSI. The above features imply that the pilot design presented in [5] (under the Rayleigh fading assumption and a total transmission power constraint) is not necessarily good for a multiuser MIMO system. New insights and guidelines must be developed to better understand the tradeoff between the channel estimation quality and the information throughput in a multiuser MIMO system.

In this paper, we investigate the fundamental throughput limit of a training-based MIMO multiple-access system, where each transmission frame consists of a training phase for acquiring CSI and a data-transmission phase for information delivery. We assume that the system consists of an N-antenna BS and K single-antenna user terminals, forming an N-by-K virtual MIMO channel with coherence time T (during which the channel state is assumed to be constant). Each user suffers both large-scale fading caused by the near-far effect and small-scale fading caused by the multipath effect. As large-
Due to the near-far effect, the distribution of the MIMO channel matrix is rotationally variant. As a result, the technique developed in [5] is not applicable to a MIMO multiuser system. Instead, we use the majorization theory [12], [19] as a basic tool to derive upper and lower bounds of the system throughput. We show that the derived throughput lower bound is asymptotically optimal for throughput maximization in the high SNR regime. We also show that the upper and lower bounds are reasonably tight in various system settings. We note that the results in this paper are applicable to an arbitrary antenna and user configuration. Further, to establish a close link with massive MIMO, we use the random matrix theory to derive a closed-form expression of the system throughput as $N, K,$ and $T$ scale at the same speed towards infinity, under the assumption of uniform large-scale fading (i.e., users are co-located).

An interesting finding of this work is that, in a training-based multiuser MIMO system, the optimal training length $\alpha T$ for throughput maximization is usually less than the number of active users $K$. This is in contrast with the case of conventional MIMO in which the optimality always occurs at $K = \alpha T$ (implying that each user has one separate time slot for channel estimation) [5]. To understand this fact, we first note that a MIMO multiuser system does not reduce to conventional MIMO even if uniform large-scale fading is assumed. The key difference is that the total transmission power of a MIMO multiuser system scales linearly with the number of active users, while that of a conventional MIMO system is usually limited by a fixed total power budget. Consequently, for a multiuser MIMO system, the scalability of the total transmission power shifts the optimality point from $K = \alpha T$ to $K \leq \alpha T$. Particularly, when the optimality occurs at $K < \alpha T$, there is not enough degrees of freedom to design orthogonal training sequences (with length $\alpha T$) for all the $K$ users. This is again different from the case of conventional MIMO in which the optimal training sequences are always orthogonal [5].

The disparity between the optimal $\alpha T$ and $K$ is further enlarged by the near-far effect. In fact, for throughput maximization in the considered multiuser MIMO system, a portion of users with relatively poor channel quality should be kept silent in transmission. This is because, with the near-far effect, the channel qualities of users vary significantly from each other. The channels of far-off users are so weak that it will be a waste of resource for throughput enhancement if any time or power is allocated to these users. As such, a good strategy is to inactivate these far-off users in transmission, which translates to an enlarged gap between the optimal $\alpha T$ and $K$.

A. Other Related Work

Existing work related to the throughput analysis of multiuser MIMO systems includes [13]–[17]. Specifically, Marzetta examined the training-throughput tradeoff in a multiuser MIMO broadcast channel by assuming orthogonal training sequences [15]. Kobayashi et. al studied the throughput optimisation of a multiuser MIMO system by taking into account the effect of channel estimation error and finite channel state feedback [16]. Chi et al. considered the pilot sequence design for a multiuser MIMO OFDM system, and derived the optimal pilot design in the sense of minimizing the mean-square error (MSE) of the channel estimation [17]. Given the above work, the characterization of the optimal training-throughput tradeoff is still missing in the literature, which motivates the work presented in this paper.

B. Organization

The remainder of this paper is organized as follows. First, in Section II we describe our system model and formulate the throughput optimization problem. Then, in Section III and Section IV we establish upper and lower bounds of the system throughput under the assumption of arbitrary large-scale fading. Later in Section V we derive the optimal design of the system parameters under the assumption of uniform large-scale fading. We conclude the paper in Section VI.

C. Notation

Bold upper-case letters denote a matrix and bold-lower case letters denote a column vector. For a matrix $H$, the element of row $i$ and column $j$ is denoted as $H_{ij}$. $I_n$ denotes an $n \times n$ identity matrix, where $n$ is an integer. The superscripts $(\cdot)^T$, $(\cdot)^\dagger$ stand for the transpose and Hermitian transpose, respectively. The operators $(\cdot)^{-1}$, $\lVert \cdot \rVert$ denote the inverse, the norm of a vector, and $\otimes$ to denote the Kronecker product. The vector inequality $x \succeq y$ means that $x$ majorizes $y$; $\log$ denotes logarithm with base 2. $(x)^+ = \max \{0, x\}$; $\text{diag}\{a_1, \ldots, a_n\}$ represents the diagonal matrix with the $(i,i)$th element being $a_i$; $(A)_{i\text{diag}}$ represents the diagonal matrix obtained by setting the off-diagonal elements of $A$ to zeros; $\text{vdiag}\{A\}$ represents the vector specified by the diagonal of $A$, with the $i$th entry of $\text{vdiag}\{A\}$ being $A_{ii}$. For a square matrix $A$, $\lambda(A)$ represents the vector of the eigenvalues of $A$ (counting multiplicity) arranged in a descending order.

II. Preliminaries

A. System Model

Consider a multiuser MIMO system, where $K$ single-antenna users deliver information to an $N$-antenna base station (BS). Assume that both $K$ and $N$ are very large but finite. The channel is block-fading, i.e., the channel keeps invariant scale fading varies much slowly in magnitude of time than small-scale fading [13], we assume that the large-scale fading coefficients of users are known a priori at the BS, while the small-scale fading coefficients of users are to be estimated using training sequences. Our target is to characterize and optimize the throughput of a training-based multiuser MIMO system over the parameters including the pilot symbols, the time allocation coefficient $\alpha$ (which specifies the fraction of the training phase in a transmission frame), the power allocation coefficient $\gamma_k$ of each user $k$, and the user number $K$, etc.
within coherence time $T$. The corresponding channel model for a frame of $T$ symbols is given by

$$Y = \sum_{k=1}^{K} d_k h_k x_k T + W,$$  \hfill (1a)

or equivalently

$$Y = HDX + W,$$  \hfill (1b)

where $Y \in \mathbb{C}^{N \times T}$ represents the received signal matrix at the BS, and $X = [x_1, \cdots, x_K]^T \in \mathbb{C}^{K \times T}$ is the transmit signal matrix with the $k$-th row $x_k^T$ being the signal of user $k$, and $H \in \mathbb{C}^{N \times K}$ is the small-scale channel fading matrix with the $(i,j)$-th element $H_{ij}$ connecting the $j$-th user to the $i$-th receive antenna of the BS, $D = \text{diag}\{d_1, d_2, \cdots, d_K\}$ with each $d_k$ being a nonnegative large-scale fading coefficient of user $k$, and $W \in \mathbb{C}^{N \times T}$ is the white Gaussian noise matrix with the power of each element given by $N_0$. Without loss of generality, we always assume that $\{d_k\}$ are arranged in a descending order, i.e., $d_1 \geq d_2 \geq \cdots \geq d_K$. The power constraint of each user $k$ is given by

$$\frac{1}{T}\|x_k\|^2 \leq P_0, \; k \in \mathcal{I}_K \triangleq \{1, 2, \cdots, K\}. \hfill (2)$$

The large-scale fading, due to signal propagation over large distances and shadowing from obstacles in the propagation path, usually varies relatively slowly. We assume that the channel matrix $H$ varies relatively slowly. We assume that the channel matrix $H$ is known a priori at the BS, as the BS is able to acquire an accurate estimate of each $d_k$ based on historical data. The small-scale fading is caused by multipath propagation. We assume that the channel matrix $H$ for small-scale fading follows independent Rayleigh fading, with each $H_{ij}$ independently drawn from $\mathcal{CN}(0, 1)$. Note that $H$ varies from frame to frame, and needs to be estimated at the BS based on the received data in each transmission frame.

When users are co-located, we have $d_1 = \cdots = d_K = d$, or equivalently, $D = dI_K$, for a certain coefficient $d$. In this case, the system in (1) is very similar to a conventional MIMO system. The only difference resides in the power constraint: In the considered multiuser MIMO system, the total transmission power is linear in the number of users, while in a conventional MIMO system, the total power constraint is usually invariant to the number of transmit antennas (i.e., the number of users in our model). In general, the system in (1) allows the existence of the near-far effect, i.e., the users are geographically separated in a random manner, thereby resulting in different values of $\{d_k\}$.

The transmission protocol for (1) is described as follows. We adopt a training-based scheme in which each transmission frame consists of two phases. In the first phase (referred to as the training phase), pilot symbols known to the receiver are transmitted, based on which the channel matrix $H$ is estimated. In the second phase (referred to as the data transmission phase), data are transmitted and detected based on the estimated channel. The details of these two phases are described below.

### B. Training Phase

Without loss of generality, we assume that $\alpha T$ channel uses are assigned to the training phase, where $\alpha \in (0, 1)$ is a coefficient to be optimized. From (1b), the channel model for the training phase is

$$Y_p = HDX_p + W_p,$$  \hfill (3)

where $X_p \in \mathbb{C}^{K \times \alpha T}$ is the pilot symbol matrix with $x_{p,k}$ being the transpose of the $k$-th row, and $W_p \in \mathbb{C}^{N \times \alpha T}$ is the corresponding AWGN. The power of user $k$ in the training phase is given by

$$\frac{1}{\alpha T}\|x_{p,k}\|^2 = \gamma_k P_0, \; k \in \mathcal{I}_K, \hfill (4)$$

where $x_{p,k}^T$ is the $k$-th row of $X_p$, and $\gamma_k$ is a power allocation coefficient of user $k$ for the training phase.

The base stations use $X_p$ and $Y_p$ to generate an estimate of the channel $H$, denoted as $\hat{H} = f(X_p, Y_p)$. Particularly, the minimum mean-square error (MMSE) estimate of $H$ is given by

$$\hat{H} = Y_p (X_p^T D^2 X_p + N_0 I_{\alpha T})^{-1} X_p^T D.$$ \hfill (5)

The corresponding MMSE matrix is given by

$$R_{\text{MMSE}} = \mathbb{E}\left[\text{vec}(H - \hat{H}) (\text{vec}(H - \hat{H}))^\dagger\right] = I_N \otimes M_{\hat{H}}, \hfill (6)$$

where vec($H - \hat{H}$) is the transpose of the row vector obtained by sequentially stacking the rows of $(H - \hat{H})$, the expectation $\mathbb{E}$ is taken over $H$ and $W$, and

$$M_{\hat{H}} = I_K - DX_p (X_p^T D^2 X_p + N_0 I_{\alpha T})^{-1} X_p^T D.$$ \hfill (7)

### C. Data Transmission Phase

In the data transmission phase, the users transmit data and the base stations carry out coherent detection based on the channel estimate obtained in the training phase. The channel model is written as

$$Y_d = \hat{H} DX_d + V,$$ \hfill (8a)

where

$$V \triangleq (H - \hat{H}) DX_d + W_d,$$ \hfill (8b)

and $X_d \in \mathbb{C}^{K \times (1-\alpha)T}$ is a zero-mean data matrix, and $W_d$ is the corresponding AWGN. The power consumption at user $k$ is expressed as

$$\frac{1}{(1-\alpha)T}\|x_{d,k}\|^2 = \gamma'_k P_0,$$ \hfill (9)

where $x_{d,k}^T$ is the $k$-th row of $X_d$, and $\gamma'_k$ is a coefficient of user $k$. With (4) and (9), the power constraint in (2) can be equivalently expressed as

$$\alpha \gamma_k + (1-\alpha) \gamma'_k \leq 1.$$

\hfill (10)
The covariance matrices of $X_d$ and $V$ are respectively given by
\[
R_{X_d} \triangleq \frac{1}{(1-\alpha)T} E \left[ X_d X_d^\dagger \right] = \text{diag}\{\gamma'_1 P_0, \ldots, \gamma'_K P_0\} \quad (11a)
\]
\[
R_V \triangleq \frac{1}{(1-\alpha)T} E \left[ V V^\dagger \right] = \frac{1}{(1-\alpha)^2} E \left[ (H-\tilde{H})D R_{X_d} D (H-\tilde{H})^\dagger \right] + N_0 I_N \quad (11b)
\]
\[
= \sigma_v^2 I_N, \quad (11c)
\]
where the equivalent noise power is given by
\[
\sigma_v^2 = \text{tr} (M_D^2 R_{X_d}) + N_0. \quad (12)
\]
In the above, (11a) follows from the fact that the user signals are independent of each other; (11b) follows by noting $V$ in (8b), and (11c) by (7) and noting that the rows of $H - \tilde{H}$ are independent of each other.

Recall the signal model in (8). The interference-plus-noise term $V$ is in general correlated with the signal $H D X_d$, which complicates the analysis. However, it is known that the “worst-case” noise for the additive channel in (8a) follows an independent Gaussian distribution [3]. That is, the instantaneous achievable rate over the channel (8a) is lower bounded by
\[
I (X_d; Y_d | \tilde{H}) = \log \left| I_N + R_{X_d} \tilde{H} D R_{X_d} \tilde{H}^\dagger \right| = \log \left| I_N + \frac{1}{\gamma} \tilde{H} D R_{X_d} \tilde{H}^\dagger \right|, \quad (13)
\]
where $I (X_d; Y_d | \tilde{H})$ is the conditional mutual information between $X_d$ and $Y_d$ calculated by assuming that the elements of the $i$th row of $X_d$ are independently drawn from $\mathcal{CN}(0, \gamma_k P_0)$ for $k = 1, \ldots, K$, and those of $V$ are independently drawn from $\mathcal{CN}(0, \sigma_v^2)$. Then, by considering the two-phase protocol and averaging over the channel fading, we obtain an achievable throughput of the system given by
\[
R = (1-\alpha) E \left[ \log \left| I_N + \frac{1}{\gamma^2} \tilde{H} D R_{X_d} \tilde{H}^\dagger \right| \right]. \quad (14)
\]

D. Problem Statement

We are interested in the throughput limit of the considered training-based multiuser MIMO system for a given number of antennas at the BS and a given coherence time, i.e., both $N$ and $T$ are fixed. With (14), the throughput maximization problem is formulated as follows:

maximize $R$ in (14) \hspace{1cm} (15a)

subject to

$$
\frac{1}{\alpha T} \|X_{p,k}\|^2 = \gamma_k P_0, \; 0 \leq \alpha \leq 1 \quad (15b)
$$

$$
\alpha \gamma_k + (1-\alpha) \gamma'_k \leq 1, \; \gamma_k \geq 0, \; k \in \mathcal{I}_K. \quad (15c)
$$

A similar problem for a training-based conventional MIMO system has been previously studied in [1], [2], [3]. Particularly, it is known that the optimal $K$ is $K = \alpha T$ for conventional MIMO [3]. In this paper, we tackle the problem in a more challenging multiuser setup. A major difference is that, due to the near far effect, the large-scale fading coefficients $\{d_k\}$ of users are in general very different from each other. In this case, it is generally difficult to find the optimal training matrix $X_p$. Instead, we derive tight upper and lower bounds of the throughput. Furthermore, we derive the optimal system design for throughput maximization under the setting of uniform large-scale fading, i.e., $D = d I_K$. Interestingly, we will show that $K = \alpha T$ is not necessarily optimal for a training-based multiuser MIMO system, even in the setting of uniform large-scale fading.

III. Throughput Upper Bound

In this section, we establish a useful throughput upper bound by relaxing the constraints of (15), as detailed below.

A. Problem Relaxation

To start with, we focus on the optimization of the pilot matrix by assuming that the other parameters $\{\gamma'_k\}, \{\gamma_k\}, K$, and $\alpha$ are given. Then, problem (15) reduces to

maximize $R$ in (14) \hspace{1cm} (16a)

subject to $\left( X_p X_p^\dagger \right)_{\text{diag}} = R X$ \hspace{1cm} (16b)

where
\[
R_X = \text{diag} \{\alpha \gamma_1 P_0 T, \ldots, \alpha \gamma_K P_0 T\}. \quad (17)
\]

The expectation in (14) is taken over $\tilde{H}$. The randomness of $\tilde{H}$ comes from the randomness of $Y_p$. We see from (3) that $Y_p$ is a zero-mean random matrix with covariance $X_p^\dagger D^2 X_p + N_0 I_{\alpha T}$. Thus, $Y_p$ can be equivalently expressed as
\[
Y_p = G \left( X_p^\dagger D^2 X_p + N_0 I_{\alpha T} \right)^{\frac{1}{2}}, \quad (18)
\]
where $G \in \mathbb{C}^{N \times T}$ is a random matrix with the elements independently drawn from $\mathcal{CN}(0, 1)$. Combining (5) and (18), we rewrite the sum rate in (16a) as
\[
R = (1-\alpha) E \left[ \log \left| I_N + \frac{1}{\gamma^2} \tilde{X} D R_{X_d} \tilde{X}^\dagger \right| \right]. \quad (19)
\]
where the expectation is taken over $G$, and
\[
\tilde{X} = DX_p (X_p^\dagger D^2 X_p + N_0 I_{\alpha T})^{-\frac{1}{2}} \quad (20)
\]
\[
\sigma_v^2 = \text{tr} \left\{ (I_K - \tilde{X} \tilde{X}^\dagger) D^2 R_{X_d} \right\} + N_0. \quad (21)
\]
Note that (21) is obtained by substituting (7) into (12).

We now consider the following problem:

maximize $R$ in (19) \hspace{1cm} (22a)

subject to $\left( \tilde{X} \tilde{X}^\dagger \right)_{\text{diag}} \leq D^2 R_X (N_0 I_K + D^2 R_X)^{-1} \hspace{1cm} (22b)$

where the relation “$\leq$” means less than or equal to in an entry-by-entry manner. The theorem below reveals that (22) is a relaxation of (16). The proof of Theorem 1 is presented in Appendix A.

Theorem 1. The optimal sum rate of problem (16) is upper bounded by the optimal sum rate of (22).
B. Pilot Design

We now present the solution to (22) that serves as a throughput upper bound of the original problem (16).

To proceed, we consider the following eigen-decomposition:

$$\tilde{X}^\dagger D_{X_p} \tilde{X} = U A U^\dagger$$  \hspace{1cm} (23)

where $U \in \mathbb{C}^{\alpha T \times \alpha T}$ is a unitary matrix, and $A = \text{diag}\{\lambda_1, \ldots, \lambda_{\min(K, \alpha T)}\} \in \mathbb{C}^{\alpha T \times \alpha T}$ is a diagonal matrix with $\lambda_i$ being the $i$th eigenvalue of $\tilde{X}^\dagger D_{X_p} \tilde{X}$. Note that $G$ is unitarily invariant since the elements of $G$ are i.i.d. Gaussian. That is, $GU$ has the same distribution as $G$ does. Therefore, the throughput in (19) can be equivalently written as

$$R = (1 - \alpha) E \left[ \log |I_N + \frac{1}{\alpha} G A G^\dagger|^2 \right].$$  \hspace{1cm} (24)

Then, problem (22) can be equivalently written as

$$\begin{align*}
\text{maximize} & \quad R \text{ in (24)} \\
\text{subject to} & \quad (R_{X_p}^2)_{\text{diag}} \leq R_X D^4 R_X (N_0 I_K + D^2 R_X)^{-1}.
\end{align*}$$  \hspace{1cm} (25)

The optimal solution to problem (25) is presented below, with the proof given in Appendix B.

**Theorem 2.** For $K \leq \alpha T$, the optimal $X_p$ to (25) satisfies

$$X_p X_p^\dagger = R_X.$$  \hspace{1cm} (26)

For $K > \alpha T$, the optimal $X_p$ to (25) satisfies the following conditions:

1. $(\tilde{X}^\dagger D^2 R_{X_p})_{\text{diag}} = R_{X_p} D^2 R_{X_p} (N_0 I_K + D^2 R_{X_p})^{-1}$;
2. $\lambda(\tilde{X}^\dagger D^2 R_{X_p})$, with the last $(K - \alpha T)$ entries being zeros, is the smallest vector that majorizes the diagonal of $R_{X_p} D^4 R_{X_p} (N_0 I_K + D^2 R_{X_p})^{-1}$.

**Remark 1.** We now describe an explicit approach to construct the optimal $X_p$ in Theorem 2. For the case of $K \leq \alpha T$, the optimal $X_p$ can be represented as $X_p = R_{X_p} \frac{1}{\sqrt{2}} U$ where $U \in \mathbb{C}^{\alpha T \times \alpha T}$ is an arbitrary orthonormal matrix satisfying $UU^\dagger = I_K$. The construction of optimal $X_p$ for the case of $K > \alpha T$ is more involving. First, we have the following equalities:

$$\begin{align*}
(\tilde{X}^\dagger D^2 R_{X_p})_{\text{diag}} &= (R_{X_p}^2 \tilde{X}^\dagger D \tilde{X} R_{X_p}^\dagger)_{\text{diag}} \quad (27a) \\
\lambda(\tilde{X}^\dagger D^2 R_{X_p}) &= \lambda(R_{X_p}^2 \tilde{X}^\dagger D \tilde{X} R_{X_p}^\dagger) \quad (27b)
\end{align*}$$

where the second equality follows by noting the fact that matrices $AB$ and $BA$ share the same set of nonzero eigenvalues. Then, from Theorem 4.3.32 in [20], for any two vectors $x > y$, a Hermitian matrix with $x$ being the eigenvalues and $y$ being the diagonal can be explicitly constructed. This, together with (27) and the two conditions in Theorem 2 ensures that the Hermitian matrix $R_{X_p}^2 \tilde{X}^\dagger D \tilde{X} R_{X_p}^\dagger$ indeed exists and can be explicitly constructed. Thus, we obtain $\tilde{X}^\dagger$. With the definition in (20), we can determine $D X_p = U_p \Sigma_p V_p$ based on the eigen-decomposition of $X \tilde{X}^\dagger$, and hence $X_p$ is constructed.

**Remark 2.** Another issue with Theorem 2 is to determine the smallest $\lambda(\tilde{X}^\dagger D^2 R_{X_p})$, with the last $(K - \alpha T)$ entries being zeros, that majorizes the diagonal of $R_{X_p} D^4 R_{X_p} (N_0 I_K + D^2 R_{X_p})^{-1}$. Without loss of generality, denote this smallest vector by $\lambda^*$. Then, the word “smallest” means that, for any vector $\lambda$ with the last $(\alpha T - K)$ entries being zeros, if $\lambda$ majorizes the diagonal of $R_{X_p} D^4 R_{X_p} (N_0 I_K + D^2 R_{X_p})^{-1}$, then $\lambda$ majorizes $\lambda^*$. The explicit construction of $\lambda$ is presented in Appendix C.

The following is an immediate result of Theorems 1 and 2.

**Corollary 1.** For $K \leq \alpha T$, the optimal $X_p$ to (16) satisfies $X_p X_p^\dagger = R_X$.

**Proof:** From Theorems 1 and 2, we see that $X_p$ satisfying $X_p X_p^\dagger = R_X$ provides a throughput upper bound for problem (16). Further, it can be readily verified that such an $X_p$ falls in the feasible region of (16). Therefore, $X_p$ satisfying $X_p X_p^\dagger = R_X$ achieves the optimum of problem (16).

**Remark 3.** For the case of $K > \alpha T$, the optimal solution to problem (22) in Theorem 2 in general only gives a throughput upper bound of problem (16). The reason is that the equality in (25) holds when achieving the optimal solution of (22), which generally goes beyond the feasible region of (16).

C. Optimization of $\{\gamma_k\}$ and $\{\gamma'_k\}$

With the pilot design in Theorem 2 we proceed to the optimization of the power coefficients $\{\gamma_k\}$ and $\{\gamma'_k\}$ for any given values of $K$ and $\alpha$.

1. The Case of $K \leq \alpha T$: For $K \leq \alpha T$, the pilot design in (26) is optimal. Then, the throughput is given by (24) with

$$\begin{align*}
\sigma_{v}^2 &= \sum_{k=1}^{K} \gamma_k^2 d_k^2 P_0 + N_0 \quad (28) \\
\lambda_k &= \frac{\alpha \gamma_k \rho_k d_k^2 T}{1 + \alpha \gamma_k \rho_k d_k^2 T}, \quad k = 1, \ldots, K \quad (29) \\
\rho_0 &= \frac{P_0}{N_0} \quad (30)
\end{align*}$$

With the above, the optimization problem can be written as

$$\begin{align*}
\text{maximize} & \quad R \text{ in (24)} \\
\text{subject to} & \quad \alpha \gamma_k + (1 - \alpha) \gamma_k' \leq 1, \\
& \quad \gamma_k \geq 0, \gamma_k' \geq 0, \quad k \in I_K.
\end{align*}$$  \hspace{1cm} (31)

We have the following two observations. First, for any given $\gamma_k \neq 0$, $R$ in (24) is monotonically increasing in $\gamma_k$. Thus, the optimal $\gamma_k$ satisfies $\alpha \gamma_k + (1 - \alpha) \gamma_k' = 1$ for any $\gamma_k \neq 0$. Second, for $\gamma_k' = 0$, $R$ is invariant to $\gamma_k$. Therefore, problem (31) reduces to

$$\begin{align*}
\text{maximize} & \quad R \text{ in (24)} \\
\text{subject to} & \quad 0 \leq \gamma_k \leq \frac{1}{\alpha}, \quad k \in I_K
\end{align*}$$  \hspace{1cm} (32)
with $\gamma_k = \frac{1-\alpha\gamma_k}{1-\alpha}$. To solve the above problem, we introduce an auxiliary variable $t$, and convert (32) to the following form:

$$\max_{\{\gamma_k\}, t} (1-\alpha)E\left[\log |I_N + \frac{1}{t}GAG^\dagger|\right] \tag{33a}$$

subject to $0 \leq \gamma_k \leq \frac{1}{\alpha}, \sigma_v^2 \leq t, k \in I_K$ \tag{33b}

where $\gamma_k = \frac{1-\alpha\gamma_k}{1-\alpha}, \sigma_v^2$ and $\lambda_k$ are respectively given by (28) and (29). By noting that $R$ in (24) is monotonically increasing in $\sigma_v^2$, we can readily show that problem (33) yields the same solution as (32).

We now show that for any given value of $t > 0$, problem (33) is a convex problem. To see this, we first note that the target function (33a) is concave and monotonically increasing and (33) is a convex problem. To see this, we first note that the convex composition rule, (33a) is a concave function of $\{\lambda_k\}$. Further, it can be readily shown that $\sigma_v^2$ is concave in $\{\gamma_k\}$. Thus, problem (33) is convex for any given value of $t$, and can be solved by convex programming together with an exhaustive search over $t > 0$.

To get more intuitions, we present an explicit solution to (32) at high SNR, with the proof given in Appendix D.

**Theorem 3.** As $\rho_0$ tends to infinity, the asymptotically optimal $\{\gamma_k\}$ and $\{\gamma_k^\dagger\}$ for problem (32) satisfy the following conditions: $\gamma_k = \gamma = \frac{1}{\alpha(1+\sqrt{(1-\alpha)^2})}$ and $\gamma_k^\dagger = 1-\alpha\gamma_k^\dagger = \frac{1-\alpha\gamma_k}{1-\alpha}$ for $k = 1, \ldots, K$.

**Remark 4.** From Theorem 3 we see that the optimal power allocation for $K \leq \alpha T$ at high SNR is to allocate an equal amount of power for channel estimation for every user, no matter how good or bad the channel of a user is. Later in Section IV we will show that this is not the case when $K > \alpha T$. In fact, a portion of weak users should be deactivated in transmission when $K > \alpha T$.

2) The Case of $K > \alpha T$: We now consider the case of $K > \alpha T$. We first show that $\sigma_v^2$ in (21) is still given by (28). To see this, we have

$$\sigma_v^2 = \sum_{k=1}^{K} \gamma_k^\dagger d_k^2 P_0 - \text{tr}\left\{XX^\dagger D^2 R_X\right\} + N_0$$

$$= \sum_{k=1}^{K} \gamma_k^\dagger d_k^2 P_0 - \text{tr}\left\{X_R X^\dagger D^2 R_X \left(N_0 I + D^2 R_X\right)^{-1}\right\} + N_0$$

$$= \sum_{k=1}^{K} \frac{\gamma_k^\dagger d_k^2 P_0}{1 + \alpha\gamma_k^\dagger P_0} + N_0$$ \tag{34}

where the first step follows from (21), and the second step from Condition 1 of Theorem 3. Correspondingly, the throughput is written as

$$R = (1-\alpha)E\left[\log |I_N + \frac{1}{\sigma_v^2} G\text{diag}\{\lambda_1, \ldots, \lambda_{\alpha T}\} G^\dagger|\right] \tag{35}$$

where $\{\lambda_k\}$ are determined by the fact that $\lambda = (\lambda_1, \ldots, \lambda_{\alpha T}, 0, \ldots, 0)$ (with the last $K - \alpha T$ entries being zeros) is the minimum vector that majorizes the diagonal of $\text{R}_X D^4 \text{R}_X \left(N_0 I + D^2 \text{R}_X\right)^{-1}$. For any given $\gamma_k^\dagger$, $R$ is non-decreasing in $\gamma_k$. Therefore, the equality $\alpha \gamma_k + (1-\alpha) \gamma_k^\dagger = 1$ holds at the maximizer.

The corresponding optimization problem is written as

$$\max_{\{\gamma_k\}, t} E\left[\log |I_N + \frac{1}{t} G\text{diag}\{\lambda_1, \ldots, \lambda_{\alpha T}\} G^\dagger|\right] \tag{36a}$$

subject to $\lambda > \text{vdiag}\left\{X_R X^\dagger D^4 X \left(N_0 I + D^2 \text{R}_X\right)^{-1}\right\}$ \tag{36b}

$$0 \leq \gamma_k \leq \frac{1}{\alpha}\gamma_k^\dagger = \frac{1-\alpha\gamma_k}{1-\alpha}, \sigma_v^2 \leq t, k \in I_K.$$ \tag{36c}

As the target function in (36a) is Schur-concave in $\{\lambda_k\}$, the optimal $\lambda$ is the minimum vector satisfying (36b). From Appendix C the optimal $\lambda$ and $\text{vdiag}\left\{X_R X^\dagger D^4 X \left(N_0 I + D^2 \text{R}_X\right)^{-1}\right\}$ are linearly related by (57). Then, it can be shown that the target function in (36a) is concave in $\{\gamma_k\}$, and that $\sigma_v^2$ is convex in $\{\gamma_k\}$. Therefore, for any given value of the auxiliary variable $t$, (36) is solvable using convex programming. Finally, the optimal solution to (36) can be found by an exhaustive search over $t$.

**D. Summary**

To summarize, the throughput upper bound developed in this section can be obtained as follows. For any given values of $\alpha$ and $K$, the optimal pilot matrix is given by Theorem 2. Then, for the case of $K \leq \alpha T$, the power allocation coefficients $\{\lambda_k\}$ and $\{\gamma_k^\dagger\}$ can be determined by solving (33) using convex programming plus a one-dimensional search; for the case of $K > \alpha T$, the power allocation coefficients $\{\gamma_k\}$ and $\{\gamma_k^\dagger\}$ can be determined by solving (36). Finally, the optimal $\alpha$ and $K$ can be found by a two-dimensional exhaustive search.

## IV. THROUGHPUT LOWER BOUND

In this section, we establish a throughput lower bound. We start with the pilot design.

### A. Pilot Design

The pilot design used in the throughput lower bound is presented in Algorithm 1. For the case of $K \leq \alpha T$, the pilot in Algorithm 1 takes the optimal form given in (25), which is the same as the case in the upper bound. The difference occurs in the case of $K > \alpha T$ where there is not enough time slots for the users to conduct orthogonal channel estimation. Intuitively, in this case, the channels of some distant users are so weak that allocating any time or power resource to these users leads to a degradation of the overall system performance. As such, a good strategy is to keep these weak users silent in transmission. Consequently, in Algorithm 1 the last $K - \alpha T$ diagonal elements of $R_X$ are set to zeros. As the diagonal elements of $D$ are arranged in a descending order, this implies that Algorithm 1 selects $\alpha T$ active users with the largest large-scale fading coefficients $\{d_k\}$ in transmission. Later, we will show that this choice is asymptotically optimal at high SNR.
Algorithm 1 Design of \( X_p \) (Throughput Lower Bound)

Input: \( K, \alpha, \gamma_1, \cdots, \gamma_K \).
Output: \( X_p \).
if \( K \leq \alpha T \) then
    Construct \( X_p \) satisfying \( X_p X_p^H = R_X \).
otherwise \{ \( K > \alpha T \) \}
    Construct \( X_p \) satisfying \( X_p X_p^H = \overline{R_X} \), where \( \overline{R_X} \) is obtained from \( R_X \) by setting its last \( K - \alpha T \) diagonal elements to zeros.
end if

B. Optimization of \( \{\gamma_k\} \) and \( \{\gamma'_k\} \)

With the pilot design in Algorithm 1 we maximize the throughput over \( \{\gamma_k\} \) and \( \{\gamma'_k\} \) as follows. For the case of \( K \leq \alpha T \), \( X_p \) in Algorithm 1 is exactly the same as the one used in the upper bound. Therefore, the optimization problem of \( \{\gamma_k\} \) and \( \{\gamma'_k\} \) is still given by \( (33) \), with the high-SNR optimal solution given by Theorem 3. For the case of \( K > \alpha T \), Algorithm 1 chooses \( \alpha T \) active users in transmission. Thus, the optimization problem is still given in the form of \( (33) \). The only difference is that the optimization is now limited to the \( K = \alpha T \) active users with the largest \( d_k \) values, with the power coefficients corresponding to the inactive users set to \( \gamma_k = \gamma'_k = 0 \). In the next subsection, we show that the above lower bound is asymptotically tight in the high SNR regime.

C. Asymptotic Analysis

In this subsection, we analyze the asymptotic behaviour of the above lower bound at high SNR. The main result is presented below, with the proof given in Appendix C.

Theorem 4. The pilot matrix \( X_p \) given by Algorithm 1 is asymptotically optimal in the sense of maximizing the throughput \( R \) in \( (15) \) as \( \rho_0 \) goes to infinity. The optimal user selection for throughput maximization at high SNR is to select \( \min\{K, \alpha T\} \) active users with the largest \( d_k \) values in transmission. The corresponding optimal \( \{\gamma_k\} \) and \( \{\gamma'_k\} \) are given by \( \gamma_k = \gamma = \frac{1}{\alpha \sqrt{1 + \frac{1}{\min\{K, \alpha T\}}} \alpha} \) and \( \gamma'_k = \frac{1 - \alpha^2}{1 - \alpha} \) if user \( k \) is active; otherwise, \( \gamma_k = \gamma'_k = 0 \).

Remark 5. The optimal power allocation derived in Theorem 4 is similar to that in Theorem 3. In fact, Theorem 4 extends the result of Theorem 3 to include the case of \( K > \alpha T \). Specifically, for \( K > \alpha T \), the optimal design is to deactivate the \( K - \alpha T \) users with the smallest large-scale fading coefficients in transmission.

D. Numerical Results

We now present numerical results to examine the tightness of the established upper and lower bounds. Fig. 1 demonstrates the throughput upper and lower bounds versus the user number \( K \) with \( \rho_0 = 40 \) dB and \( \rho_0 = 50 \) dB. The upper bound is calculated based on the result in Section III. The lower bound is given by Algorithm 1. The simulation settings are as follows: \( N = 100, T = 200 \). The large-scale fading coefficient \( d_k \) of each user \( k \) is modelled as \( d_k = r_k^{-2} \) (corresponding to a large-scale fading exponent of 4), where \( r_k \) represents the distance between the base station and user \( k \). It is assumed that users are uniformly distributed in a circle with radius 100 meters. Also, in simulation, the optimal \( \alpha \) is obtained by exhaustive search for each given value of \( K \). From Fig. 1 we see that the upper and lower bounds are very tight under various values of \( \rho_0 \) and \( K \).

Fig. 2 also includes the throughput behaviour for random pilots. For random pilots, every element of the pilot matrix \( X_p \) is randomly drawn from a complex normal distribution \( CN(0, 1) \); then each row \( k \) of \( X_p \) is scaled to meet the power constraint of user \( k \). The power allocation coefficients are set to \( \gamma_1 = \cdots = \gamma_K = \gamma'_1 = \cdots = \gamma'_K = 1 \). We see that the proposed optimal pilot design significantly outperforms the random pilot design, especially in the case of \( \rho_0 = 50 \) dB.

Fig. 2 illustrates the throughput lower bound and the throughput for random pilots against the SNR \( \rho_0 \) with \( K = 100 \) and \( K = 120 \).
under various values of SNR, while the opposite is observed for the random pilot design. We also see that at high SNR, the lower bound outperforms the random pilot design in the medium to high SNR regime. We also see that the proposed lower bound significantly performs better when the number of users increases from 60 to 80, while the opposite is observed for the random pilot design. Specifically, we set $\rho = 80$ dB, $\rho = 60$ dB, and $\rho = 40$ dB. The simulation settings follow those of Section V. We see that the optimal number of active users increases with SNR $\rho_0$.

Fig. 3. The throughput against the SNR $\rho_0$ with various values of $N$. 

Fig. 4. The optimal number of active users against the total user number $K$ under various values of SNR $\rho_0$. 

60 and $K = 80$. The simulation settings follow those of Fig. 1. We see that the proposed lower bound significantly outperforms the random pilot design in the medium to high SNR regime. We also see that at high SNR, the lower bound performs better when the number of users increases from 60 to 80, while the opposite is observed for the random pilot design. The reason is as follows. For the lower bound, increasing the number of users provides more freedom to select the set of active users with better channels; however, for random pilot design, as all the users are active, more users imply higher interference.

Fig. 3 is similar to Fig. 2 but with different settings. Specifically, we set $K = 120$ and $T = 200$ and $N = 100, 150$. Again, we see that the proposed lower bound significantly outperforms the random pilot design. Also, we see that the throughput grows with the increase of $N$, due to the beamforming gain of the BS antenna array.

Fig. 4 illustrates the optimal number of active users against the number of users $K$ with SNR $\rho_0 = 40, 60, 80$ dB. The other settings follow Fig. 1. In Fig. 4, the optimal number of active users may be significantly less than the number of available users $K$. We see that the optimal number of active users increase monotonically with $K$. Intuitively, the reason is that, with a larger $K$, there are more users with relatively good channels to be activated for transmission. Moreover, we also see from Fig. 4 that the optimal number of active users increase with SNR $\rho_0$.

V. THROUGHPUT OPTIMIZATION WITH UNIFORM LARGE-SCALE FADING

In this section, we consider the throughput maximization when all the users are co-located, i.e., $D = dI_K$. We derive the optimal training design for this setup, and compare the optimal design with the upper and lower bounds in the preceding sections. Due to user symmetry, we always assume a common power allocation factor $\gamma$, i.e., $\gamma = \gamma_1 = \cdots = \gamma_K$.

A. Optimal Pilot Design

The following theorem gives the optimal pilot design under the assumption of uniform large-scale fading, with the proof presented in Appendix F.

**Theorem 5.** Assume $d_1 = d_2 = \cdots = d_K = d$ and $\gamma_1 = \gamma_2 = \cdots = \gamma_K = \gamma$. Then, for $K \leq \alpha T$, the optimal training matrix $X_p$ to (16) satisfies

$$X_p X_p^\dagger = \alpha \gamma P_0 T I_K;$$

(37a)

For $K > \alpha T$, the optimal training matrix satisfies

$$X_p^0 X_p = \gamma P_0 K I_{\alpha T}$$

(37b)

$$||x_{p,k}||^2 = \alpha \gamma P_0 T, \quad k \in I_K.$$  
(37c)

**Remark 6.** An explicit construction of the optimal pilot matrix $X_p$ in Theorem 5 is described as follows. We focus on the case of $K > \alpha T$, as the case of $K \leq \alpha T$ is straightforward. For $K > \alpha T$, $X_p \in \mathbb{C}^{K \times \alpha T}$ is a tall matrix. To meet the conditions in (37b) and (37c) simultaneously, $X_p$ can be formed by extracting $\alpha T$ columns of the $K$-by-$K$ normalized discrete-Fourier-transform (or Hadamard) matrix.

**Remark 7.** An implication of Theorem 5 is that, if the optimal $K$ occurs at $K > \alpha T$, then the optimal training sequences (i.e., the rows of $X_p$) are not orthogonal to each other. Later, we will see that the optimal $K$ may occur at $K > \alpha T$ for the multiuser MIMO system in consideration. This is in contrast with the case of conventional MIMO where the optimal training sequences are always orthogonal.

**Remark 8.** It is also interesting to compare the optimal pilot design in Theorem 5 with the upper bound in Theorem 2. For the case of $K \leq \alpha T$, it can be readily shown that (26) reduces to (37a) by letting $\gamma_1 = \gamma_2 = \cdots = \gamma_K = \gamma$. Thus, both theorems give the same pilot design for $K \leq \alpha T$. What is more interesting is the case of $K > \alpha T$. In this case, it can be shown that the pilot design satisfying (37b) and (37c) in general does not meet the two conditions specified in Theorem 2.
Therefore, Theorem 2 only provides an upper bound even in the case of uniform large-scale fading, i.e., $d_1 = d_2 = \cdots = d_K = d$.

Based on Theorem 5, we simplify the throughput expression as follows. First note that $\sigma_2^2$ in (21) can be rewritten as
\[
\sigma_2^2 = \begin{cases} 
\gamma' d^2 P_0 N_0 K, & K \leq \alpha T \\
\alpha d^2 P_0 (K-\alpha T) + N_0, & K > \alpha T \end{cases} \tag{38a}
\]
where
\[
\gamma' = \frac{1-\alpha}{1-\alpha}. \tag{39}
\]
It is worth noting that, for the upper bound in Theorem 2, the expression of $\sigma_2^2$ is given by (38a) for both $K \leq \alpha T$ and $K > \alpha T$.

With the above, the throughput in (19) can be expressed as
\[
R = (1-\alpha) \mathbb{E} \left[ \log |I_N + \tau GG^\dagger| \right] \tag{40}
\]
where the elements of $G \in \mathbb{C}^{N \times \min\{K, \alpha T\}}$ are independently drawn from $\mathcal{CN}(0, 1)$, and
\[
\tau = \begin{cases} 
\frac{\alpha \gamma' \rho_0^2 d^4 T}{\rho_0 d^2 (K + \alpha T) + 1}, & K \leq \alpha T \\
\frac{\alpha \gamma' \rho_0^2 d^4 (K-\alpha T) + \rho_0 d^2 K (\gamma' + 1) + 1}{\gamma' \rho_0^2 d^4 (K-\alpha T) + \rho_0 d^2 K (\gamma' + 1) + 1}, & K > \alpha T \end{cases} \tag{41a}
\]
with SNR $\rho_0 = \frac{\rho_0}{N_0}$. Then, the optimization problem can be rewritten as
\[
\begin{align*}
\text{maximize} & \quad R \text{ in } \tag{42a} \\
\text{subject to} & \quad 0 \leq \alpha \leq 1, \quad 0 \leq \gamma \leq \frac{1}{\alpha}, \tag{42b}
\end{align*}
\]
B. Optimization over $\gamma$

We first optimize the power coefficient $\gamma$. For any fixed values of $K$ and $\alpha$, the coefficient $\gamma$ is only related to $\tau$ in (40). Thus, the optimization problem with respect to $\gamma$ can be written as
\[
\begin{align*}
\text{maximize} & \quad \tau \text{ in } \tag{43a} \\
\text{subject to} & \quad 0 \leq \gamma \leq \frac{1}{\alpha}, \tag{43b}
\end{align*}
\]
where $\tau$ is defined below (40). The solution to (43) is presented below.

**Theorem 6.** The optimal $\gamma$ to (43) is given by
\[
\gamma_{\text{opt}} = \begin{cases} 
\frac{1}{\alpha(1 + \sqrt{1 - \mu_1})}, & K \leq \alpha T \\
1, & K > \alpha T \end{cases} \tag{44a}
\]
where $\mu_1 = \rho_0 d^2 (K-1)/(1-\alpha + \rho_0 d^2 K)$ and $\mu_2 = \rho_0 d^2 (2\alpha-1)/(1-\alpha + \rho_0 d^2 K)$.

**Proof:** The optimal $\gamma$ for (43) is readily obtained by solving the Karush-Kuhn-Tucker (KKT) conditions (21).

**Remark 9.** We note that the optimal $\gamma$ given by (44) is continuous at $K = \alpha T$.

C. Optimization over $K$

We now consider the optimization of $K$ for given $\alpha$. We have the following main result, with the proof given in Appendix C.

**Theorem 7.** For any given $\alpha \in [0, 1]$, the optimal $K = K_{\text{opt}}$ for (43) is given by
\[
K_{\text{opt}} = \max \left\{ \frac{x^*}{\rho_0 d^2}, \alpha T \right\}, \tag{45}
\]
where $x^*$ is the root of
\[
f(x) = -x^2 - x \sqrt{\alpha - \alpha^2} \left( \frac{x^* + \alpha}{x + 1 - \alpha} + \frac{x^* + 1 - \alpha}{x + \alpha} \right) + 2 \sqrt{\alpha - \alpha^2}(x + \alpha)(x + 1 - \alpha) + 2(\alpha - \alpha^2). \tag{46}
\]

For a general SNR $\rho_0$, we have no explicit expression of the optimal $K$ in terms of $\alpha$. But we have a closed-form expression of the optimal $K$ in the high SNR regime, as presented below.

**Corollary 2.** For any given $\alpha \in [0, 1]$, the optimal $K$ to (43) satisfies
\[
K \to \alpha T, \quad \text{as } \rho_0 \to \infty. \tag{47}
\]

**Proof:** The corollary holds by letting $\rho_0 \to \infty$ in (45).

**Remark 10.** The tradeoff involved in optimizing $K$ are elaborated as follows. On one hand, for the data transmission phase, it is known from the information theory that the throughput of the considered multiuser MIMO channel increases unboundedly as the user number $K$ tends to infinity, provided that the user channels are perfectly known. On the other hand, for the training phase, the channel estimation accuracy decreases in a growing $K$, as more channel coefficients need to be estimated as $K$ increases. From Theorem 7, the optimal tradeoff occurs when $K = \alpha T$. Further, Corollary 2 states that the optimal $K$ tends to $\alpha T$ in the high SNR regime.

**Remark 11.** The remaining issue is to optimize $\alpha$. Though it is difficult to derive an explicit expression, the optimal $\alpha$ can be readily obtained by using an exhaustive search over $[0, 1]$.

D. Large System Analysis

In this subsection, we present an approximate expression of the throughput by using the random matrix theory. Let $\lambda$ be a non-zero eigenvalue of $\frac{1}{N} GG^\dagger$. Then

**Proposition 1.** As $N, K, T \to \infty$ with fixed ratios of $K/N = \beta$ and $K/T = \omega$, the asymptotic distribution of $\lambda$ is given by
\[
f_{\beta, \omega}(\lambda) = \begin{cases} 
\frac{\sqrt{(\lambda - \beta)^2 + (b - \lambda)^2}}{2\lambda}, & \omega \leq \alpha \tag{48a} \\
\frac{\sqrt{(\lambda - \beta)^2 + (b - \lambda)^2}}{2\lambda}, & \omega > \alpha \tag{48b}
\end{cases}
\]
where $a = (1 - \sqrt{\beta})^2$, $b = (1 + \sqrt{\beta})^2$, $a' = (1 - \sqrt{\alpha \beta / \omega})^2$ and $b' = (1 + \sqrt{\alpha / \omega})^2$.

**Proof:** Recall that the elements of $G \in \mathbb{C}^{N \times \min\{K, \alpha T\}}$ are independently drawn from $\mathcal{CN}(0, 1)$. Then, the proof is immediate by noting Theorem 2.35 in [18].
With Proposition 1, the throughput in (40) can be approximated by (49). It is known that the large-system approximation is accurate even when the system parameters $N, K, T$ are relatively small [18]. Compared with (40), an advantage of (49) is that no Monte Carlo simulation is required in evaluating the throughput. Also, (49) provides a simpler analytical throughput characterization when applying the analytical methods in this work to a massive MIMO setup.

![Fig. 5. The comparison of the real throughput and its asymptotic expression.](image)

**E. Numerical Results**

We now present numerical results to verify the analysis. Fig. 5 illustrates the throughputs given by (49) and (40) against $K$ with various values of $\alpha$. The settings are $\gamma = 1$, $\beta = K/N = 1$, $\rho_0 = -18$ dB, $N = 100$, and $T = 200$. In Fig. 5 we see that the two curves always coincide with each other. This demonstrates that (49) is indeed a good approximation of the throughput in (40).

Fig. 6 illustrates the optimal user number $K$ versus $\alpha$ under various SNR values. The simulation settings are as follows: $N = 300$, $T = 200$, $d_1 = \cdots = d_K = d = 1$. We see that the optimal $K$ is always no less than $\alpha T$, and it converges to $\alpha T$ for an arbitrary value of $\alpha$ when $\rho_0$ goes to infinity. This is in well agreement with Theorem 5 and Corollary 2.

Fig. 7 illustrates the throughput against the SNR $\rho$ with $T = 200$ and $N = 300, 500$. The optimal pilot design and the random pilot design are considered in simulation. The optimal pilot design is given by Theorem 5. For random pilot design, we set $\gamma = \gamma' = 1$ and $K = \alpha T$; each element of the pilot matrix $X_p$ is independently drawn from a complex Gaussian distribution, and then each row of $X_p$ is scaled to meet the power constraint for each user. In both designs, the optimal $\alpha$ is obtained through an exhaustive one-dimensional search. We see that the optimal pilot design significantly outperforms the random pilot design especially in the medium to high SNR regime. We see that the throughput increases with $N$. This is because a larger $N$ implies a higher power gain of the receiving antenna array. Moreover, we also include the upper and lower bounds developed in Sections III and IV. We see that these bounds are very tight in the medium to high SNR regime.

![Fig. 6. The optimal number of users $K$ vs the time allocation factor $\alpha$ under various values of SNR $\rho_0$.](image)

**VI. Conclusions**

In this paper, we analyzed the tradeoff between the channel estimation quality and the information throughput of a training-based multiuser MIMO system. We first studied the pilot design, and established upper and lower bounds for throughput maximization. We then considered the optimization of power coefficients for training and data transmission. We showed that the established lower bound is asymptotically optimal in the high SNR regime.

Our analysis revealed that the optimal training design strategies for a multiuser MIMO system are very different from those for conventional MIMO. For example, we showed that
the optimal training sequences for throughput maximization in a multiuser MIMO system are not necessarily orthogonal to each other. Also, due to the near-far effect, the optimal training design strategy for throughput maximization is to inactivate a portion of users with the weakest channels in transmission. These findings will provide insights and guidance for the practical design of a training-based multiuser system.

Future research may arise in a number of directions. For example, the work in this paper was focused on the case of a single BS with multiple antennas, without considering the interference from other base stations. BS cooperation can significantly mitigate such interference; see, e.g., [22–24] and the references therein. How to characterize the tradeoff between the channel estimation overhead and the information throughput in a cellular network with BS cooperation will be a challenging research topic deserving future endeavor.

APPENDIX A
PROOF OF THEOREM 1

As problems (16) and (22) have the same objective function, it suffices to show that the feasible region of (22) contains that of (16), or in other words, to show that (22b) is implied by (16b). To proceed, let the singular value decomposition of $DX_p$ be

$$DX_p = U_p \Sigma_p V_p^t$$

where $\Sigma_p \in \mathbb{C}^{K \times \alpha T}$ is a diagonal matrix with non-negative diagonal elements, and $U_p \in \mathbb{C}^{K \times K}$ and $V_p \in \mathbb{C}^{\alpha T \times \alpha T}$ are unitary matrices. Then, (16b) can be equivalently expressed as

$$\sum_{j=1}^{K} |u_{k,j}|^2 \sigma_{p,j}^2 \leq \frac{\sum_{j=1}^{K} |u_{k,j}|^2 \sigma_{p,j}^2}{N_0 + \alpha d_2 \gamma_k P_0 T}, \quad k \in I_K$$

where $u_{k,j}$ is the $(j,k)$th element of $U_p$; $\sigma_{p,j}$ is the $j$th diagonal element of $\Sigma_p$ for $1 \leq j \leq \min(K, \alpha T)$ and $\sigma_{p,j} = 0$ otherwise. Note that $\sum_{j=1}^{K} |u_{k,j}|^2 = 1$, and that $\frac{\sigma_{p,j}^2}{\sqrt{N_0 + \alpha d_2 \gamma_k P_0 T}}$ is a concave function of $\sigma_{p,j}$. Thus, we obtain

$$\sum_{j=1}^{K} |u_{k,j}|^2 \sigma_{p,j}^2 \leq \frac{\sum_{j=1}^{K} |u_{k,j}|^2 \sigma_{p,j}^2}{N_0 + \alpha d_2 \gamma_k P_0 T} = \frac{\alpha d_2 \gamma_k P_0 T}{N_0 + \alpha d_2 \gamma_k P_0 T}, \quad k \in I_K$$

where the first step is from the Jensen’s inequality, and the second step from (51). Noting that (52) is equivalent to (22b), we conclude the proof of Theorem 1.

APPENDIX B
PROOF OF THEOREM 2

For ease of discussion, we first ignore the effect of $\sigma_{p,j}^2$ in throughput maximization. The minimization of $\sigma_{p,j}^2$ (so as to maximize the throughput) is discussed at the end of the proof.

From Theorem 4.3.26 in [20], we obtain

$$\lambda(R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}^\frac{1}{2}) > \text{vdiag}(R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}^\frac{1}{2})$$

The definition of the majorization “≻” can be found in Appendix C. From the matrix theory, $R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}$ and $\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}$ share the same set of nonzero eigenvalues. That is, $\lambda(R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d})$ and $\lambda$ have the same set of nonzero elements. Thus, problem (22) can be recast as to maximize the rate in (24) subject to (25b) and (53). As the rate in (24) is monotonically increasing in each $\lambda_j$, we see that the maximum of the problem occurs when (25b) takes the equality. Therefore, problem (22) can be rewritten as

maximize $R$ in (24) subject to

$$\lambda \succ \text{vdiag}(R_{X_d}D^4R_X(N_0I_K + D^2R_X)^{-1})$$

where $\lambda = (\lambda_1, \cdots, \lambda_K)$ for $K \leq \alpha T$, and $\lambda = (\lambda_1, \cdots, \lambda_{\alpha T}, 0, \cdots, 0)$ for $K > \alpha T$.

We now solve problem (53) for two different cases, namely, $K \leq \alpha T$ and $K > \alpha T$. For the case of $K \leq \alpha T$, we have $\lambda = (\lambda_1, \cdots, \lambda_K)$. Note that $R$ in (24) is symmetric and concave with respect to $\lambda_1, \cdots, \lambda_K$. From Proposition C.2 in [19], $R$ in (24) is Schur concave. Thus, the optimal $\lambda$ to problem (53) is given by $\lambda = \text{vdiag}(R_{X_d}D^4R_X(N_0I_K + D^2R_X)^{-1})$. This implies that the optimal $X_p$ to (22) satisfies (26), which proves the first half of the theorem.

For the case of $K > \alpha T$, the rank of $R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}$ is at most $\alpha T$, we have $\lambda = (\lambda_1, \cdots, \lambda_{\alpha T}, 0, \cdots, 0) \in \mathbb{C}^K$. Thus, unlike the case of $K \leq \alpha T$, we cannot set $\lambda$ to $\text{vdiag}(R_{X_d}D^4R_X(N_0I_K + D^2R_X)^{-1})$. Then, from the Schur convexity of the throughput function, the optimal $\lambda$ should be the smallest vector that majorizes $\text{vdiag}(R_{X_d}D^4R_X(N_0I_K + D^2R_X)^{-1})$. Noting the equivalence between problems (22) and (54), we see that the optimal $X_p$ to (22) satisfies the two conditions in the second half of the theorem.

To complete the proof of Theorem 2, we still need to show that the above pilot design of $X_p$ also minimizes the equivalent noise power $\sigma_{n}^2$ in (21) (and hence indeed maximizes the throughput). From (21), we see that minimizing $\sigma_{n}^2$ is equivalent to maximizing

$$\text{tr}(\bar{X}\bar{X}^tD^2R_X) = \text{tr}(R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}^\frac{1}{2})$$

where $\bar{X}$ is defined in (20). Clearly, $\text{tr}(R_{X_d}^\frac{1}{2}D\bar{X}\bar{X}^t\bar{D}^\frac{1}{2}R_{X_d}^\frac{1}{2})$ is maximized when the equality in (25b) holds. This agrees with the optimality condition (26) for the case of $K \leq \alpha T$; it also agrees with condition 1 in Theorem 2 for the case of...
Therefore, we obtain by definition that increasing rearrangements of $\frac{1}{j-m} \sum_{i=m+1}^{j} x[i] \leq \frac{1}{k-m} \sum_{i=m+1}^{k} x[i] \leq \frac{1}{k-m} \sum_{i=m+1}^{k} y[i]$, $j = m + 1, \ldots, k$ \hspace{1cm} (60)

where the first inequality follows from the fact that $x[i_{m+1}], x[i_{m+2}], \ldots, x[i_k]$ are in an increasing order, and the second inequality follows from $x \succ y$. Together with (57b), we have

$$\sum_{i=m+1}^{j} x[i] \leq \frac{j-m}{k-m} \sum_{i=m+1}^{k} y[i] = \sum_{i=m+1}^{j} x[*]_j, \quad j = m + 1, \ldots, k.$$ \hspace{1cm} (61)

The proof concludes by combining (59) and (61).

### APPENDIX D

**Proof of Theorem 3**

Let $\hat{G}$ be the submatrix of $G$ that consists of the first $K$ columns of $G$. Then, $R = (1 - \alpha) \log \left| I_N + \hat{G} \text{diag}(\frac{1}{\gamma_1}, \ldots, \frac{1}{\gamma_K}) \hat{G}^\dagger \right|$ As $\rho_0 \to \infty$, 

$$\log \left| I_N + \hat{G} \text{diag}(\frac{1}{\gamma_1}, \ldots, \frac{1}{\gamma_K}) \hat{G}^\dagger \right|$$

can be approximated by

$$\log \left( \prod_{k=1}^{K} \lambda_k \right) - K \log \sigma_n^2 + \log \left| \hat{G}^\dagger \hat{G} \right|.$$ 

Then, the throughput can be written as

$$R = (1 - \alpha) \sum_{k=1}^{K} \log \frac{1 - \alpha \gamma_k}{\sum_{k=1}^{K} (1 - \frac{1}{\gamma_k}) + 1 - \alpha} + (1 - \alpha) \log \left| \hat{G}^\dagger \hat{G} \right|. \hspace{1cm} (62)$$

Then, the optimization problem reduces to

$$\text{maximize } R \text{ in (62)} \hspace{1cm} (63a)$$

subject to $0 \leq \gamma_k \leq \frac{1}{\alpha}, \quad k \in I_K.$ \hspace{1cm} (63b)

The solution to problem (63) is described as follows. Note that

$$\log(1 - \alpha \gamma_k)$$

is concave in $\gamma_k$, and that

$$\log(\frac{1}{\alpha} \sum_{k=1}^{K} (1 - \frac{1}{\gamma_k}) + 1 - \alpha)$$

is convex in $\gamma_k$. Thus, $R$ in (63) is concave in $\{\gamma_k\}$, and so problem (63) is a convex problem. By solving the KKT conditions, we obtain $\gamma_1 = \cdots = \gamma_K = \gamma$ and $\gamma'_1 = \cdots = \gamma'_K = \gamma'$, which concludes the proof.

### APPENDIX E

**Proof of Theorem 4**

From Corollary 1, we see that Algorithm 1 outputs the optimal $X_p$ in the sense of throughput maximization when $K \leq \alpha T$. The corresponding optimal $\{\gamma_k\}$ and $\{\gamma'_k\}$ at high SNR are given by Theorem 3, which is in agreement with the statement in Theorem 4. Then, it suffices to focus on the case of $K > \alpha T$, i.e., $X_p$ is a tall matrix.

Without loss of generality, we assume that there are $K'$ active users with $K' \leq K$. Let $X_p \in \mathbb{C}^{K' \times \alpha T}$ be the pilot matrix for those active users, and $D \in \mathbb{C}^{K' \times K'}$ be the diagonal fading-coefficient matrix. We consider the limiting process of $X_p = \sqrt{\rho_0} W_p$ as $\rho_0 \to \infty$, where $W_p$ is an arbitrary constant invariant to $\rho_0$.

An explicit construction of $X_p$ satisfying conditions 1 and 2 is presented in Remarks 1 and 2.

2 For an inactive user $k$, we have $\gamma_k = \gamma'_k = 0$. 

# APPENDIX C

**Backgrounds of the Majorization Theory**

In this appendix, we introduce some background knowledge of the majorization theory [19] used in this paper. For $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, $x \succ y$ if

$$\sum_{i=1}^{k} x[i] \leq \sum_{i=1}^{k} y[i], \quad k = 1, \ldots, n - 1$$

Further, we obtain

$$\sum_{i=m+1}^{j} x[i] \leq \frac{j-m}{k-m} \sum_{i=m+1}^{k} y[i] = \sum_{i=m+1}^{j} x[*]_j, \quad j = m + 1, \ldots, k.$$ \hspace{1cm} (61)

The proof concludes by combining (59) and (61).

**Lemma 1.** $x^* \succ y$.

**Proof:** We first note that $x^*$ defined in (57), together with $S$ in (58), is already arranged in an increasing order. Further, it is clear that

$$\sum_{i=1}^{k} x[i] \leq \sum_{i=1}^{k} y[i], \quad k = 1, \ldots, n - 1$$

Therefore, we obtain by definition that $x^* \succ y$.

**Lemma 2.** Let $x$ be an $n$-by-1 nonnegative vector containing $m$ zeroes. Then, $x \succ y$ implies $x \succ x^*$.

**Proof:** From $x \succ y$ and (57), we obtain

$$\sum_{i=1}^{j} x[i] = \sum_{i=1}^{j} x[i] = 0, \quad j = 1, \ldots, t$$

Further, we obtain

$$\sum_{i=1}^{j} x[i] \leq \sum_{i=1}^{j} y[i] = \sum_{i=1}^{j} x[*]_j, \quad j = k, \ldots, n - 1.$$ \hspace{1cm} (59b)
Let the compact SVD of $\overline{D}W_p$ be $\overline{D}W_p = \overline{U}_p\Sigma_p\overline{V}_p^\dagger$, where $\overline{U}_p \in \mathbb{C}^{K \times r}$ satisfies $\overline{U}_p^\dagger\overline{U}_p = I_r$, $\Sigma_p \in \mathbb{C}^{r \times r}$ is a diagonal matrix with the diagonal elements being the singular values, $\overline{V}_p \in \mathbb{C}^{C \times r}$ satisfies $\overline{V}_p^\dagger\overline{V}_p = I_r$, and $r \leq \alpha T$ is the rank of $\overline{W}_p$. Then, as $\rho_0 \to \infty$, we obtain

$$\overline{D}X_p(\overline{X}_pD\overline{X}_p^\dagger + N_0I_{\alpha T})^{-\frac{1}{2}}\overline{D}$$

(64a)

$$= \rho_0\overline{D}W_p(\rho_0\overline{W}_pD\overline{W}_p + N_0I_{\alpha T})^{-\frac{1}{2}}\overline{D}W_p$$

(64b)

$$= \rho_0\overline{U}_p\Sigma_p^2(\rho_0\overline{U}_p\Sigma_p^2 + N_0I_r)^{-1}\Sigma_p\overline{U}_p$$

(64c)

$$= \overline{U}_p\Sigma_p^2$$

(64d)

Thus, as $\rho_0 \to \infty$, the equivalent noise power $\sigma_v^2$ in (21) satisfies

$$\sigma_v^2 \to \text{tr}\left\{(\overline{I}_{K'} - \overline{U}_p\overline{U}_p^\dagger)\overline{D}R_{X_d}\right\} + N_0$$

(65)

where $R_{X_d} \in \mathbb{C}^{K' \times K'}$ is the diagonal matrix obtained by deleting the rows and columns of $R_{X_d}$ corresponding to inactive users. Note that $I_{K'} - \overline{U}_p\overline{U}_p^\dagger \neq 0$ for $K' > \alpha T$, i.e., $\sigma_v^2$ in (65) is unbounded when $K' > \alpha T$. This implies that the optimal number of active users for throughput maximization cannot exceed $\alpha T$. Moreover, from Theorem 3 to maximize throughput in the case of $K' \leq \alpha T$, all the $K'$ users should be active. Therefore, the optimal number of active users for throughput maximization is given by $K' = \alpha T$.

We now discuss how to select the $\alpha T$ active users. Without loss of generality, we assume that users 1, $\cdots$, $\alpha T$ are the selected active users. This implies that $\gamma_{\alpha T + 1} = \cdots = \gamma_K = \gamma_{\alpha T + 1} = \cdots = \gamma_K = 0$. Then, as $\rho_0 \to \infty$, the throughput in (24) can be approximated by $\log \prod_{k=1}^{\alpha T} \left(\frac{1}{\alpha T}\right) + \log |\overline{G}| |\overline{G}|$, where $\sigma_v^2$ is given by (28), $\lambda_k$ is given by (29) for $k = 1, \cdots, \alpha T$, and $\overline{G} \in \mathbb{C}^{N \times \alpha T}$ consists of the first $\alpha T$ columns of $G$. Further note that, to maximize $\prod_{k=1}^{\alpha T} \left(\frac{1}{\alpha T}\right)$ at high SNR, $\alpha \lambda_k + (1 - \alpha)\frac{1}{\alpha T} = 1$ for $k = 1, \cdots, \alpha T$. Then the throughput is given by (62) with $K$ replaced by $\alpha T$. From (62), we see that the optimal choice to maximize the throughput is to select users with the maximum $d_k$ values. Finally, the optimal $\{\gamma_k\}$ for the active users are given by Theorem 3 which concludes the proof.

APPENDIX F
PROOF OF THEOREM 7

Following the proof of Theorem 2, we first ignore the effect of $\sigma_v^2$ in throughput maximization. We will discuss the minimization of $\sigma_v^2$ (so as to maximize the throughput) at the end of the proof.

With $d_1 = d_2 = \cdots = d_K = d$ and $\gamma_1 = \gamma_2 = \cdots = \gamma_K = \gamma$, we can rewrite (19) as

$$R = (1 - \alpha)\mathbb{E}\left[\log |I_N + \frac{\gamma d^2 P_0}{\sigma_v^2}G\overline{X}^\dagger\overline{X}G\dagger|\right]$$

(66)

where $\gamma' = \frac{1 - \alpha}{1 - \alpha}$, and

$$\overline{X} = dX_p(dX_p\overline{X}_p + N_0I_{\alpha T})^{-\frac{1}{2}}$$

(67)

$$= U_p\Sigma_p^2(\Sigma_p^2 + N_0I_{\alpha T})^{-\frac{1}{2}}V_p^\dagger$$

(68)

where the last step follows from the singular value decomposition: $dX_p = U_p\Sigma_p^2 V_p^\dagger$. Then, as $G\overline{V}_p$ has the same distribution as $G$ does, we can further rewrite (66) as

$$R = (1 - \alpha)\mathbb{E}\left[\log |I_N + \frac{\gamma' d^2 P_0}{\sigma_v^2}G\overline{X}^\dagger\overline{X}G\dagger|\right]$$

(69)

where $\Lambda \in \mathbb{C}^{\alpha T \times \alpha T}$ is given by

$$\Lambda = \gamma' d^2 P_0(\Sigma_p^2 + N_0I_{\alpha T})\frac{1}{\sigma_v^2}$$

(70)

$$= \text{diag}\left\{\gamma' d^2 P_0\sigma_1^2, \cdots, \gamma' d^2 P_0\sigma_{\min(K,\alpha T)}^2\right\}$$

(71)

Also, the eigenvalue vector of $d^2X_p\overline{X}_p^\dagger$ can be expressed as $x = (\sigma_1^2, \cdots, \sigma_{\min(K,\alpha T)}^2)$. From Proposition 3.26 in [20], we have $x \gg \text{vdiag}(d^2X_p\overline{X}_p^\dagger)$, where $1$ is an all-one vector of an appropriate size. Thus, problem (15) can be recast as

maximize $R$ in (69)

(70a)

subject to $x \gg \alpha \gamma' d^2 P_0T$.

(70b)

Note that $R$ in (24) is symmetric and concave with respect to $\sigma_v^2, \cdots, \sigma_{\min(K,\alpha T)}^2$. From Proposition C.2 in [19], $R$ in (24) is Schur concave. Thus, the optimal $x$ should be the smallest vector that majorizes $\alpha \gamma' d^2 P_0T$. More specifically, for the case of $K \leq \alpha T$, zero padding in $x$ is not necessary. Thus, the optimal $x$ is simply taken as $x = \alpha \gamma' d^2 P_0T$. This implies that $X_p = x = \alpha \gamma' d^2 P_0T$. For the case of $K > \alpha T$, as the rank of $\Lambda$ is at most $\alpha T$, we see that $x$ is padded with at least $K - \alpha T$ zeros. Then, the optimal $x$ is given by $x_1 = \cdots = x_{\alpha T} = \gamma' d^2 P_0$ and $x_{\alpha T + 1} = \cdots = x_K = 0$. This implies that (77) holds.

To complete the proof, we still need to show that the above pilot design of $X_p$ minimizes the equivalent noise power $\sigma_v^2$ in (21). The detailed argument is very similar to the corresponding part of the proof of Theorem 2 in Appendix E. We omit the details for brevity. This concludes the proof of Theorem 7.
Then
\[
\tau K = \frac{\alpha_0^2 d^4 K T}{\rho_0 d^2 (\gamma^{-1} K + \alpha \gamma^{-1} T) + \gamma^{-1} \gamma^{-1}} = \frac{\rho_0^2 d^4 K T}{(1-\alpha) \left( \sqrt{1 + \rho_0^2 d^2 T} + \sqrt{1 + \frac{\rho_0^2 d^2 K}{1-\alpha}} \right)^2} \tag{71}
\]
and
\[
\frac{\partial (\tau K)}{\partial K} = \frac{\rho_0^2 d^4 T}{(1-\alpha) \left( \sqrt{1 + \rho_0^2 d^2 T} + \sqrt{1 + \frac{\rho_0^2 d^2 K}{1-\alpha}} \right)^3} > 0. \tag{72}
\]

Therefore, \( \tau K \) is monotonically increasing in \( K \) for \( K \leq \alpha T \).

We now consider Fact 2. Denote
\[
g(x_1, \ldots, x_K) = \mathbb{E} \left[ \log \left[ I_N + \frac{\theta}{K} \mathbf{G} \mathbf{G}^\dagger \right] \right]. \tag{73}
\]
Clearly, \( g(x_1, \ldots, x_K) \) is symmetric and concave with respect to \( x_1, \ldots, x_K \). Thus, from Proposition C.2 in [19], \( g(x_1, \ldots, x_K) \) is Schur-concave in \( x_1, \ldots, x_K \). Let \( \mathbf{G}' \in \mathbb{C}^{N \times \min(K',\alpha T)} \) with the elements independently drawn from \( \mathcal{CN}(0,1) \). Then, we obtain
\[
\mathbb{E} \left[ \log \left[ I_N + \frac{\theta}{K} \mathbf{G} \mathbf{G}^\dagger \right] \right] = g(0/K, \ldots, 0/K) \geq g(\theta/K', \ldots, \theta/K', 0, \ldots, 0) = \mathbb{E} \left[ \log \left[ I_N + \frac{\theta}{K'} \mathbf{G} \mathbf{G}^\dagger \right] \right],
\]
where the inequality holds from the Schur-concavity of the \( g \)-function and the fact that
\[
(\theta/K', \ldots, \theta/K', 0, \ldots, 0) > (\theta/K, \ldots, \theta/K) \tag{74}
\]
for any integer \( K' < K \). Therefore, Fact 2 holds.

What remains is to determine the optimal \( K \) to maximize \( R \) in (40) for \( K \geq \alpha T \). In this case, \( R \) in (40) depends on \( K \) only through \( T \). Thus, it suffices to find the optimal \( K \) to maximize \( T \).

To proceed, we note that for \( K \geq \alpha T \), \( \gamma^{-1} = \alpha(1 + \sqrt{1-\mu_2}) \) and \( \gamma^{-1} = (1-\alpha) \left( \frac{\gamma^{-1}}{\sqrt{1-\mu_2}} + 1 \right) \) and \( \gamma^{-1} = \alpha \gamma^{-1} + (1-\alpha) \gamma^{-1} \), where \( \mu_2 = \frac{(1-\alpha)(\rho_0 d^2 K + \alpha)}{\alpha \rho_0 d^2 K + 1-\alpha} \). Then
\[
\tau = \frac{\rho_0^2 d^4 T}{\rho_0^2 d^4 K (K - \alpha T)} + \frac{\rho_0^2 d^4 K (\gamma^{-1} + \gamma^{-1}) + \gamma^{-1} \gamma^{-1}}{\rho_0^2 d^4 K N} \tag{75}
\]
where
\[
t(K) = \sqrt{(1-\alpha)(\rho_0 d^2 K + \alpha) + \sqrt{\alpha(\rho_0 d^2 K + 1 - \alpha)}}. \tag{76}
\]
It is not difficult to see that the above \( \tau \), as a function of \( K \), has exactly one extremal point (which is the maximum) in the range of \( K \geq \alpha T \). Taking derivative of \( \tau \) with respect to \( K \) and setting it to zero, we see that the optimal \( K \) satisfies \( f(\rho_0 d^2 K) = 0 \) with \( f(x) \) defined in (46). This concludes the proof of Theorem [7].