How Much Training is Needed in Multiple-Antenna Wireless Links?

Babak Hassibi and Bertrand M. Hochwald

***Abstract*—Multiple-antenna wireless communication links promise very high data rates with low error probabilities, especially when the wireless channel response is known at the receiver. In practice, knowledge of the channel is often obtained by sending known training symbols to the receiver. We show how training affects the capacity of a fading channel—too little training and the channel is improperly learned, too much training and there is no time left for data transmission before the channel changes. We compute a lower bound on the capacity of a channel that is learned by training, and maximize the bound as a function of the received signal-to-noise ratio (SNR), fading coherence time, and number of transmitter antennas. When the training and data powers are allowed to vary, we show that the optimal number of training symbols is equal to the number of transmit antennas—this number is also the smallest training interval length that guarantees meaningful estimates of the channel matrix. When the training and data powers are instead required to be equal, the optimal number of symbols may be larger than the number of antennas. We show that training-based schemes can be optimal at high SNR, but suboptimal at low SNR.**

***Index Terms*—BLAST, high-rate wireless communications, receive diversity, space–time coding, transmit diversity.**

I. Introduction

MULTIPLE-ANTENNA wireless communication links promise very high data rates with low error probabilities, especially when the wireless channel response is known at the receiver [1], [2]. To learn the channel, the receiver often requires the transmitter to send known training signals during some portion of the transmission interval. An early study of the effect of training on a multiantenna channel capacity is [3], where it is shown that, under certain conditions, by choosing the number of transmit antennas to maximize the throughput in a wireless channel, one generally spends half the coherence interval training. We, however, address a different problem: given a multiantenna wireless link with M transmit antennas, N receive antennas, coherence interval of length T (in symbols), and a signal-to-noise ratio (SNR) ρ, *how much of the coherence interval should be spent training?*

Our solution is based on a lower bound on the information-theoretic capacity achievable with training-based schemes. An example of a training-based scheme that has attracted recent attention is BLAST [4], [5], where an experimental prototype has achieved 20-b/s/Hz data rates with eight transmit and twelve receive antennas. The lower bound allows us to compute the optimal amount of training as a function of ρ, T, M, and N. We are also able to identify some occasions where training imposes a substantial information-theoretic penalty, especially when the coherence interval T is only slightly larger than the number of transmit antennas M, or when the SNR is low. In these regimes, training to learn the entire channel matrix is highly suboptimal. Conversely, if the SNR is high and T is much larger than M, then training-based schemes can come very close to achieving capacity.

We show that if optimization over the training and data powers is allowed, then the optimal number of training symbols is always equal to the number of transmit antennas. If the training and data powers are instead required to be equal, then the optimal number of symbols can be larger than the number of antennas. The reader can get a sample of the results given in this paper by glancing at the figures in Section IV. These figures present a capacity lower bound (that is sometimes tight) and the optimum training intervals as a function of the number of transmit antennas M, receive antennas N, the fading coherence time T and SNR ρ.

II. Channel Model and Problem Statement

We assume that the channel obeys the simple discrete-time *block-fading* law, where the channel is constant for some discrete time interval T, after which it changes to an independent value that it holds for another interval T, and so on. This is an appropriate model for time-division multiple access (TDMA) or frequency-hopping systems, and is a tractable approximation of a continuously fading channel model such as Jakes' [6]. We further assume that channel estimation (via training) and data transmission is to be done within the interval T, after which new training allows us to estimate the channel for the next T symbols, and so on.

Within one block of T symbols, the multiple-antenna model is

$$X = \sqrt{\frac{\rho}{M}}\, SH + V \tag{1}$$

where X is a $T \times N$ received complex signal matrix, the dimension N representing the number of receive antennas. The transmitted signal is S, a $T \times M$ complex matrix where M is the number of transmit antennas. The $M \times N$ matrix H represents the channel connecting the M transmit to the N receive

Manuscript received May 18, 2000; revised February 9, 2002. The material in this paper was presented at the 34th Asilomar Conference on Signals, Systems and Computers, Asilomar, CA, October 2000.

B. Hassibi was with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA. He is now with the California Institute of Technology, Pasadena, CA 91125 USA (e-mail: hassibi@caltech.edu).

B. M. Hochwald is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: hochwald@bell-labs.com).

Communicated by M. L. Honig, Associate Editor for Communications.
Digital Object Identifier 10.1109/TIT.2003.809594

antennas, and V is a $T \times N$ matrix of additive noise. The matrices H and V both comprise independent random variables whose mean square is unity. We also assume that the entries of the transmitted signal S have, on the average, unit mean square (see, e.g., (4)). Thus, ρ is the expected received SNR at each receive antenna. We let the additive noise V have zero-mean unit-variance independent complex-Gaussian entries. Although we often also assume that the entries of H are also zero-mean complex-Gaussian distributed, many of our results do not require this assumption.

A. Training-Based Schemes

Since H is not known to the receiver, training-based schemes dedicate part of the transmitted matrix S to be a known training signal from which we can learn H. In particular, training-based schemes are composed of the following two phases.

1) **Training Phase:** Here we may write

$$X_\tau = \sqrt{\frac{\rho_\tau}{M}}\, S_\tau H + V_\tau, \qquad S_\tau \in \mathcal{C}^{T_\tau \times M},\ \mathrm{tr}\, S_\tau S_\tau^* = MT_\tau \tag{2}$$

where S_τ is the matrix of training symbols sent over T_τ time samples and known to the receiver, ρ_τ is the SNR during the training phase, and $X_\tau \in \mathcal{C}^{T_\tau \times N}$ is the received matrix. (We allow for different transmit powers during the training and data transmission phases.) Because S_τ is fixed and known, there is no expectation in the normalization of (2).

2) **Data Transmission Phase:** Here we may write

$$X_d = \sqrt{\frac{\rho_d}{M}}\, S_d H + V_d, \qquad S_d \in \mathcal{C}^{T_d \times M},\ \mathrm{E}\,\mathrm{tr}\, S_d S_d^* = MT_d \tag{3}$$

where S_d is the matrix of data symbols sent over T_d time samples, ρ_d is the SNR during the data transmission phase, and $X_d \in \mathcal{C}^{T_d \times N}$ is the received matrix. Because S_d is random and unknown, the normalization in (3) has an expectation.

This two-part training and data process is equivalent to partitioning the matrices in (1) as

$$S = \begin{pmatrix} \sqrt{\frac{\rho_\tau}{\rho}}\, S_\tau \\ \sqrt{\frac{\rho_d}{\rho}}\, S_d \end{pmatrix}, \quad X = \begin{pmatrix} X_\tau \\ X_d \end{pmatrix}, \quad V = \begin{pmatrix} V_\tau \\ V_d \end{pmatrix}. \tag{4}$$

Conservation of time and energy yield

$$T = T_\tau + T_d, \qquad \rho T = \rho_\tau T_\tau + \rho_d T_d. \tag{5}$$

We establish a lower bound on the capacity of encoding/decoding rules that use a transmitted signal S partitioned as in (4). One possible receiver uses S_τ and X_τ to generate an estimate of the channel

$$\hat{H} = f(X_\tau, S_\tau). \tag{6}$$

Two examples include the maximum-likelihood (ML) and linear minimum mean-square error (LMMSE) estimates

$$\hat{H} = \sqrt{\frac{M}{\rho_\tau}}\, (S_\tau^* S_\tau)^{-1} S_\tau^* X_\tau$$
$$\hat{H} = \sqrt{\frac{M}{\rho_\tau}} \left(\frac{M}{\rho_\tau} I_M + S_\tau^* S_\tau\right)^{-1} S_\tau^* X_\tau. \tag{7}$$

(To obtain a meaningful estimate of H, we need at least as many measurements as unknowns, which implies that $N \cdot T_\tau \geqslant N \cdot M$ or $T_\tau \geqslant M$.) In many training-based receivers, the channel estimate is then used as if it were the true channel, and data is sent over S_d. Other conceivable receivers might process X_τ and X_d jointly with knowledge of S_τ to estimate S_d without explicitly forming $\hat{H}$.

We cannot say whether any particular transmitter/receiver structure (with a partitioned S) can achieve the bound we compute, but we can say that there exists some structure whose performance is at least as good as our bound. Receivers that assume the channel estimate after training to be perfect are generally suboptimal (sometimes also called "mismatched") and their analyses can be complicated [7], [8]. We do not pursue such analyses here; our bound does not directly apply to the best mismatched receiver—instead, our bound applies to the optimal transmitter/receiver combination. Such a receiver would, for example, exploit any statistical dependence between the channel estimation error and the data signal. However, it is reasonable to expect that at high SNR, when the channel estimate after training is accurate, our bound accurately portrays the best performance achievable by receivers that assume the channel estimate to be perfect.

Whether or not an explicit or implicit $\hat{H}$ is formed, it is clear that increasing T_τ improves the quality of $\hat{H}$, but if T_τ is too large, then $T_d = T - T_\tau$ is small and too little time is set aside for data transmission. Similarly, dedicating too much power ρ_τ to training compromises the power available to transmit data. We compute the ρ_τ and T_τ that optimize our capacity bound.

III. Capacity and Capacity Bounds

The capacity in bits per channel use is the maximum over the distribution of the transmit signal S_d of the mutual information between the known and observed signals X_τ, S_τ, X_d and the unknown transmitted signal S_d. This is written as

$$C_\tau = \sup_{p_{S_d}(\cdot),\, \mathrm{E}\|S_d\|_F^2 \leqslant MT_d} \frac{1}{T} I(X_\tau, S_\tau, X_d; S_d).$$

Now

$$I(X_\tau, S_\tau, X_d; S_d) = I(X_d; S_d | X_\tau, S_\tau) + \underbrace{I(X_\tau, S_\tau; S_d)}_{=0}$$
$$= I(X_d; S_d | X_\tau, S_\tau),$$

where $I(X_\tau, S_\tau; S_d) = 0$ because S_d is independent of S_τ and X_τ. Thus, the capacity is the supremum (over the distribution of S_d) of the mutual information between the transmitted S_d and received X_d, given the transmitted and received training signals S_τ and X_τ

$$C_\tau = \sup_{p_{S_d}(\cdot),\, \mathrm{E}\|S_d\|_F^2 \leqslant MT_d} \frac{1}{T} I(X_d; S_d | X_\tau, S_\tau). \tag{8}$$

The capacity depends on the conditional distribution of H given S_τ and X_τ.

For receiver structures that form an explicit $\hat{H}$, as long as information is not "thrown away" in the process, it is possible to achieve C_τ as given in (8). However, some data transmission schemes that employ training do throw away information because they form an explicit $\hat{H}$ and use it as if it were correct.

Our method for finding a lower bound for C_τ computes an explicit $\hat{H}$, relegates the estimation error of this channel estimate to the additive noise, and then considers only the correlation (and not the full statistical dependence) between the resulting noise and the transmitted signal. We then obtain a lower bound by replacing the resulting noise by a worst case (but analytically tractable) noise with this same correlation.

We assume that $\hat{H}$ is the conditional mean of H (which is the minimum mean-square error (MMSE) estimate), given S_τ and X_τ. During the data transmission phase, we may then write

$$X_d = \sqrt{\frac{\rho_d}{M}}\, S_d \hat{H} + \underbrace{\sqrt{\frac{\rho_d}{M}}\, S_d \tilde{H} + V_d}_{V_d'}, \tag{9}$$

where $\tilde{H} = H - \hat{H}$ is the zero-mean channel estimation error, and V_d' combines the additive noise and residual channel estimation error. By well-known properties of the conditional mean, $\hat{H}$ and $\tilde{H}$ are uncorrelated. The estimate $\hat{H}$ is known to the receiver and assumed by our lower bound computation to be correct; hence, the channel capacity of a training-based system is lower-bounded by the capacity of a *known channel* system, subject to additive noise with the power constraint

$$\begin{aligned}
\sigma_{V'}^2 &= \frac{1}{NT_d} \operatorname{tr} \mathrm{E}\, V_d' V_d'^* \\
&= \frac{1}{NT_d} \mathrm{E} \operatorname{tr} \left[\frac{\rho_d}{M} \tilde{H}\tilde{H}^* S_d^* S_d \right] + \frac{1}{NT_d} \mathrm{E} \operatorname{tr} V_d V_d^* \\
&= \frac{\rho_d}{MNT_d} \operatorname{tr} \left[\mathrm{E}(\tilde{H}\tilde{H}^*) \mathrm{E}(S_d^* S_d) \right] + 1.
\end{aligned} \tag{10}$$

There are two important differences between (9) and (1). In (9), the channel is known to the receiver whereas in (1) it is not. In (1), the additive noise is Gaussian and independent of the data whereas in (9) it is possibly neither. Finding the capacity of the known-channel system requires us to examine the worst effect the additive noise can have during data transmission. We therefore wish to find

$$\begin{aligned}
C_\tau \geqslant C_{\text{worst}} = &\inf_{p_{V_d'}(\cdot),\, \operatorname{tr} \mathrm{E}\, V_d' V_d'^* = NT_d} \\
&\cdot \sup_{p_{S_d}(\cdot),\, \operatorname{tr} \mathrm{E}\, S_d S_d'^* = MT_d} I(X_d; S_d | \hat{H}).
\end{aligned}$$

A similar argument for lower-bounding the mutual information in a scalar multiple-access wireless channel is given in [9]. The worst case noise is the content of the next theorem, which is proven in the Appendix.

Theorem 1 (Worst Case Uncorrelated Additive Noise): Consider the matrix-valued additive noise known channel

$$X = \sqrt{\frac{\rho}{M}}\, SH + V$$

where $H \in \mathcal{C}^{M\times N}$ is the known channel, and where the signal $S \in \mathcal{C}^{1\times M}$ and the additive noise $V \in \mathcal{C}^{1\times N}$ satisfy the power constraints

$$\mathrm{E}\, \frac{1}{M} SS^* = 1 \quad \text{and} \quad \mathrm{E}\, \frac{1}{N} VV^* = 1$$

and are uncorrelated

$$\mathrm{E}\, S^* V = 0_{M\times N}.$$

Let $R_V = \mathrm{E}\, V^* V$ and $R_S = \mathrm{E}\, S^* S$. Then the worst case noise has a zero-mean Gaussian distribution $V \sim \mathcal{CN}(0, R_{V,\text{opt}})$, where $R_{V,\text{opt}}$ is the minimizing noise covariance in

$$\begin{aligned}
C_{\text{worst}} = &\min_{R_V,\, \operatorname{tr} R_V = N}\ \max_{R_S,\, \operatorname{tr} R_S = M} \mathrm{E} \log \\
&\cdot \det \left(I_N + \frac{\rho}{M} R_V^{-1} H^* R_S H \right).
\end{aligned} \tag{11}$$

We also have the minimax property

$$\begin{aligned}
&I_{V\sim\mathcal{CN}(0,\, R_{V,\text{opt}}),\, S}(X; S) \\
&\quad \leqslant I_{V\sim\mathcal{CN}(0,\, R_{V,\text{opt}}),\, S\sim\mathcal{CN}(0,\, R_{S,\text{opt}})}(X; S) \\
&\quad = C_{\text{worst}} \leqslant I_{V,\, S\sim\mathcal{CN}(0,\, R_{S,\text{opt}})}(X; S)
\end{aligned} \tag{12}$$

where $R_{S,\text{opt}}$ is the maximizing signal covariance matrix in (11). When the distribution on H is left rotationally invariant, i.e., when $p(\Theta H) = p(H)$ for all Θ such that $\Theta\Theta^* = \Theta^*\Theta = I_M$, then

$$R_{S,\text{opt}} = I_M.$$

When the distribution on H is right rotationally invariant, i.e., when $p(H\Theta) = p(H)$ for all Θ such that $\Theta\Theta^* = \Theta^*\Theta = I_N$, then

$$R_{V,\text{opt}} = I_N.$$

The notion that Gaussian additive noise is the worst for mutual information is not new [10]–[12]. Theorem 1 is, however, tailored for our purposes since the noise is uncorrelated with the signal (rather than independent as is usually assumed in these references), and we are also able to compute the optimal R_S and R_V.

In our case, the additive noise and signal are uncorrelated when the channel estimate is the MMSE estimate

$$\hat{H} = \mathrm{E}_{|X_\tau,\, S_\tau} H,$$

because

$$\begin{aligned}
\mathrm{E}_{|X_\tau,\, S_\tau} S_d V_d'^* &= \mathrm{E}_{|X_\tau,\, S_\tau} S_d \left(\sqrt{\frac{\rho_d}{M}}\, S_d^* \tilde{H}^* + V_d^* \right) \\
&= \sqrt{\frac{\rho_d}{M}}\, \mathrm{E}_{|X_\tau,\, S_\tau} S_d S_d^* \tilde{H}^* + \mathrm{E}_{|X_\tau,\, S_\tau} S_d V_d^* \\
&= \sqrt{\frac{\rho_d}{M}}\, \mathrm{E}_{|X_\tau,\, S_\tau} S_d S_d^* \mathrm{E}_{|X_\tau,\, S_\tau} \tilde{H}^* + 0 \\
&= 0, \qquad \text{since } \mathrm{E}_{|X_\tau,\, S_\tau}(H - \hat{H}) = 0.
\end{aligned}$$

The MMSE estimate is the only estimate with this property.

The noise term V_d' in (9), when $\hat{H}$ is the MMSE estimate, is uncorrelated with S_d but is not necessarily Gaussian. Theorem 1 says that a lower bound on the training-based capacity is obtained by replacing V_d' by independent zero-mean additive Gaussian noise with the same power constraint $\operatorname{tr} R_{V,\,\mathrm{opt}} = N\sigma^2_{V'}$. Because $\mathrm{E}\, S_d^* S_d = T_d R_S$, (10) becomes

$$\begin{aligned}\sigma^2_{V'} &= 1 + \frac{\rho_d}{MNT_d} \operatorname{tr}\left[(\mathrm{E}\,\tilde{H}\tilde{H}^*) T_d R_S\right] \\ &= 1 + \rho_d \sigma^2_{\tilde{H},\,R_S} \end{aligned} \tag{13}$$

where $\sigma^2_{\tilde{H},\,R_S} \triangleq \frac{1}{NM}\,\mathrm{E}\operatorname{tr}\tilde{H}^* R_S \tilde{H}$. Using (11), we may, therefore, write

$$\begin{aligned} C_\tau &\geqslant C_{\mathrm{worst}} \\ &= \min_{R_V,\, \operatorname{tr} R_V = N}\; \max_{R_S,\, \operatorname{tr} R_S = M} \mathrm{E}\, \frac{T-T_\tau}{T} \log \\ &\quad \cdot \det\left(I_N + \frac{\rho_d}{1+\rho_d\sigma^2_{\tilde{H},\,R_S}}\, \frac{R_V^{-1}\hat{H}^* R_S \hat{H}}{M}\right) \end{aligned}$$

where the coefficient $T - T_\tau$ reflects the fact that the data transmission phase has a duration of $T_d = T - T_\tau$ time symbols. Since $\hat{H}$ is zero mean, its variance can be defined as $\sigma^2_{\hat{H}} = \frac{1}{NM}\,\mathrm{E}\operatorname{tr}\hat{H}^*\hat{H}$. By the orthogonality principle for MMSE estimates

$$\sigma^2_{\hat{H}} = 1 - \sigma^2_{\tilde{H}} \tag{14}$$

where $\sigma^2_{\tilde{H}} = \frac{1}{NM}\,\mathrm{E}\operatorname{tr}\tilde{H}^*\tilde{H}$. Define the *normalized channel estimate* as

$$\overline{H} \triangleq \frac{1}{\sigma_{\hat{H}}}\,\hat{H}.$$

We may write the capacity bound as

$$\begin{aligned} C_\tau &\geqslant \min_{R_V,\, \operatorname{tr} R_V = N}\; \max_{R_S,\, \operatorname{tr} R_S = M} \mathrm{E}\, \frac{T-T_\tau}{T} \log \\ &\quad \cdot \det\left(I_N + \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H},\,R_S}}\, \frac{R_V^{-1}\overline{H}^* R_S \overline{H}}{M}\right). \end{aligned} \tag{15}$$

The ratio

$$\rho_{\mathrm{eff}} = \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H},\,R_S}} \tag{16}$$

can, therefore, be considered as an *effective* SNR. This bound does not require H to be Gaussian.

The remainder of this paper is concerned with maximizing this lower bound. We consider choosing the following:

1) the training data S_τ;
2) the training power ρ_τ;
3) the training interval length T_τ.

This is, in general, a formidable task since computing the conditional mean for a channel H with an arbitrary distribution can itself be difficult. However, when the elements of H are independent $\mathcal{CN}(0,\,1)$ then the computations become manageable. In fact, in this case we have

$$\operatorname{vec}\hat{H} = R_{HX_\tau} R_{X_\tau}^{-1} (\operatorname{vec} X_\tau),$$

where

$$R_{HX_\tau} = \mathrm{E}(\operatorname{vec} H)(\operatorname{vec} X_\tau)^*$$

and

$$R_{X_\tau} = \mathrm{E}(\operatorname{vec} X_\tau)(\operatorname{vec} X_\tau)^*.$$

(The $\operatorname{vec}(\cdot)$ operator stacks all of the columns of its arguments into one long column; the above estimate of H can be rearranged to coincide with the LMMSE estimate given in (7).) Moreover, the distribution of $X_\tau = \sqrt{\frac{\rho_\tau}{M}}\, S_\tau H + V_\tau$ is rotationally invariant from the right ($p(X_\tau\Theta) = p(X_\tau)$, for all unitary Θ) since the same is true of H and V. This implies that $\hat{H}$ and $\overline{H}$, are rotationally invariant from the right. Therefore, applying Theorem 1 yields $R_{V,\,\mathrm{opt}} = I_N$.

The choice of R_S that maximizes the lower bound (15) depends on the distribution of $\overline{H}$ which, in turn, depends on the training signal S_τ. Thus, in principle, one needs to perform a joint optimization over R_S and S_τ. But we are interested in designing S_τ, and hence we turn the problem around by arguing that the optimal S_τ depends on R_S. That is, the choice of training signal depends on how the antennas are to be used during data transmission, which is perhaps more natural to specify first. We specify that the antennas are to be used such that $R_S = I_M$, which is the same as saying that we are using them independently and with equal power. This choice is reasonable because the transmitter does not know the channel, and it allows us to obtain a valid and tractable lower bound on capacity. In fact, Theorem 1 (see also [1]) says that $R_S = I_M$ is best when the distribution of $\overline{H}$ is left rotationally invariant. Section III-A shows that the choice of S_τ that maximizes ρ_{eff} gives $\overline{H}$ this property. Thus, even though we cannot claim joint optimality over R_S and S_τ, we can claim that our choice of training signal and R_S are consistent. With $R_S = I_M$, we have

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T} \log\det\left(I_N + \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H}}}\,\frac{\overline{H}^*\overline{H}}{M}\right). \tag{17}$$

A. Optimizing Over S_τ

The first parameter over which we can optimize the capacity bound is the choice of the training signal S_τ. From (17), it is clear that S_τ primarily affects the capacity bound through the effective SNR ρ_{eff}. Thus, we propose to choose S_τ to maximize ρ_{eff}[1]

$$\rho_{\mathrm{eff}} = \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H}}} = \frac{\rho_d(1-\sigma^2_{\tilde{H}})}{1+\rho_d\sigma^2_{\tilde{H}}} = \frac{1+\rho_d}{1+\rho_d\sigma^2_{\tilde{H}}} - 1.$$

It, therefore, follows that we need to choose S_τ to minimize the mean-square error $\sigma^2_{\tilde{H}}$.

Because $\sigma^2_{\tilde{H}} = \frac{1}{NM}\operatorname{tr} R_{\tilde{H}}$, we compute the covariance matrix $R_{\tilde{H}} \triangleq \mathrm{E}(\operatorname{vec}\tilde{H})(\operatorname{vec}\tilde{H})^*$ of the MMSE estimate (which in this case is also the LMMSE estimate)

$$\begin{aligned} R_{\tilde{H}} &= R_H - R_{HX_\tau} R_{X_\tau}^{-1} R_{X_\tau H} \\ &= I_M \otimes I_N - \left(\sqrt{\frac{\rho_\tau}{M}}\, S_\tau^* \otimes I_N\right) \\ &\quad \cdot \left(I_M \otimes I_N + S_\tau \frac{\rho_\tau}{M} S_\tau^* \otimes I_N\right)^{-1} \left(S_\tau\sqrt{\frac{\rho_\tau}{M}} \otimes I_N\right) \\ &= \left(I_M + \frac{\rho_\tau}{M} S_\tau^* S_\tau\right)^{-1} \otimes I_N \end{aligned}$$

[1] Maximizing SNRs has been studied in many other contexts as well; for a study in intersymbol interference (ISI) channels see [21].

where we have used the equation $X_\tau = \sqrt{\frac{\rho_\tau}{M}}\, S_\tau H + V_\tau$ to compute R_{HX_τ}, R_{X_τ}, and $R_{X_\tau H}$. It follows that we need to choose S_τ to solve

$$\min_{S_\tau,\, \operatorname{tr} S_\tau^* S_\tau = MT_\tau} \frac{1}{M} \operatorname{tr}\left(I_M + \frac{\rho_\tau}{M} S_\tau^* S_\tau\right)^{-1}.$$

In terms of $\lambda_1, \ldots, \lambda_M$, the eigenvalues of $S_\tau^* S_\tau$, this minimization can be written as

$$\min_{\substack{\lambda_1, \ldots, \lambda_M \\ \sum \lambda_m \leqslant MT_\tau}} \frac{1}{M} \sum_{m=1}^{M} \frac{1}{1 + \frac{\rho_\tau}{M}\lambda_m}$$

which is solved by setting $\lambda_1 = \cdots = \lambda_M = T_\tau$. This yields

$$S_\tau^* S_\tau = T_\tau I_M \tag{18}$$

as the optimal solution; i.e., *the training signal must be a multiple of a matrix with orthonormal columns*. A similar conclusion is drawn in [3] when training for BLAST and [13] when training with so-called “shift-invariant” sequences to minimize total estimation error.

With this choice of training signal, we obtain

$$\sigma_{\tilde{H}}^2 = \frac{1}{1 + \frac{\rho_\tau}{M} T_\tau} \quad \text{and} \quad \sigma_{\hat{H}}^2 = \frac{\frac{\rho_\tau}{M} T_\tau}{1 + \frac{\rho_\tau}{M} T_\tau}. \tag{19}$$

In fact, we have the stronger result

$$R_{\tilde{H}} = \frac{1}{1 + \frac{\rho_\tau}{M} T_\tau} I_M \otimes I_N$$

and

$$R_{\hat{H}} = \frac{\frac{\rho_\tau}{M} T_\tau}{1 + \frac{\rho_\tau}{M} T_\tau} I_M \otimes I_N \tag{20}$$

which implies that $\overline{H} = \frac{1}{\sigma_{\hat{H}}} \hat{H}$ has independent $\mathcal{CN}(0, 1)$ entries, and is, therefore, rotationally invariant.

Thus, (17) can be written as

$$C_\tau \geqslant \mathrm{E}\, \frac{T - T_\tau}{T} \log\det\left(I_M + \rho_{\text{eff}} \frac{\overline{H}\,\overline{H}^*}{M}\right) \tag{21}$$

where

$$\rho_{\text{eff}} = \frac{\rho_d \rho_\tau T_\tau}{M(1 + \rho_d) + \rho_\tau T_\tau} \tag{22}$$

and where $\overline{H}$ has independent $\mathcal{CN}(0, 1)$ entries.

B. Optimizing Over the Power Allocation

Recall that the effective SNR is given by

$$\rho_{\text{eff}} = \frac{\rho_d \rho_\tau T_\tau}{M(1 + \rho_d) + \rho_\tau T_\tau}$$

and that the power allocation $\{\rho_d, \rho_\tau\}$ enters the capacity formula via ρ_{eff} only. Thus, we need to choose $\{\rho_d, \rho_\tau\}$ to maximize ρ_{eff}. To facilitate the presentation, let α denote the fraction of the total transmit energy that is devoted to the data

$$\rho_d T_d = \alpha \rho T, \quad \rho_\tau T_\tau = (1 - \alpha)\rho T, \qquad 0 < \alpha < 1. \tag{23}$$

Therefore, we may write

$$\begin{aligned} \rho_{\text{eff}} &= \frac{\rho_d \rho_\tau T_\tau}{M(1 + \rho_d) + \rho_\tau T_\tau} = \frac{\alpha \frac{\rho T}{T_d} \cdot (1 - \alpha)\rho T}{M\left(1 + \alpha \frac{\rho T}{T_d}\right) + (1 - \alpha)\rho T} \\ &= \frac{(\rho T)^2}{T_d} \cdot \frac{\alpha(1 - \alpha)}{M + \rho T - \rho T\left(1 - \frac{M}{T_d}\right)\alpha} \\ &= \frac{\rho T}{T_d - M} \cdot \frac{\alpha(1 - \alpha)}{-\alpha + \frac{M + \rho T}{\rho T\left(1 - \frac{M}{T_d}\right)}}. \end{aligned}$$

To maximize ρ_{eff} over $0 < \alpha < 1$ we consider the following three cases.

1) $T_d = M$:

$$\rho_{\text{eff}} = \frac{(\rho T)^2}{M(M + \rho T)} \alpha(1 - \alpha).$$

It readily follows that

$$\alpha = \tfrac{1}{2} \tag{24}$$

and, therefore, that

$$\rho_d = \frac{T}{2M}\rho, \; \rho_\tau = \frac{T}{2(T - M)}\rho, \; \rho_{\text{eff}} = \frac{(\rho T)^2}{4M(M + \rho T)}.$$

2) $T_d > M$: We write

$$\rho_{\text{eff}} = \frac{\rho T}{T_d - M} \cdot \frac{\alpha(1 - \alpha)}{-\alpha + \gamma}, \quad \gamma = \frac{M + \rho T}{\rho T\left(1 - \frac{M}{T_d}\right)} > 1.$$

Differentiating and noting that $\gamma > 1$ yields

$$\arg\max_{0 < \alpha < 1} \frac{\alpha(1 - \alpha)}{-\alpha + \gamma} = \gamma - \sqrt{\gamma(\gamma - 1)},$$

from which it follows that

$$\rho_{\text{eff}} = \frac{\rho T}{T_d - M}\left(\sqrt{\gamma} - \sqrt{\gamma - 1}\right)^2. \tag{25}$$

3) $T_d < M$: We write

$$\rho_{\text{eff}} = \frac{\rho T}{M - T_d} \cdot \frac{\alpha(1 - \alpha)}{\alpha - \gamma}, \quad \gamma = \frac{M + \rho T}{\rho T\left(1 - \frac{M}{T_d}\right)} < 0.$$

Differentiating and noting that $\gamma < 0$ yields

$$\arg\max_{0 < \alpha < 1} \frac{\alpha(1 - \alpha)}{\alpha - \gamma} = \gamma + \sqrt{\gamma(\gamma - 1)}$$

from which it follows that

$$\rho_{\text{eff}} = \frac{\rho T}{M - T_d}\left(\sqrt{-\gamma} - \sqrt{-\gamma + 1}\right)^2. \tag{26}$$

We summarize these results in a theorem.

Theorem 2 (Optimal Power Distribution): The optimal power allocation $\alpha = \frac{\rho_d T_d}{\rho T}$ in a training-based scheme is given by

$$\alpha = \begin{cases} \gamma - \sqrt{\gamma(\gamma - 1)}, & \text{for } T_d > M \\ \frac{1}{2}, & \text{for } T_d = M \\ \gamma + \sqrt{\gamma(\gamma - 1)}, & \text{for } T_d < M \end{cases} \tag{27}$$

where $\gamma = \frac{M + \rho T}{\rho T\left(1 - \frac{M}{T_d}\right)}$. The corresponding capacity lower bound is

$$C_\tau \geqslant \mathrm{E}\, \frac{T - T_\tau}{T} \log\det\left(I_M + \rho_{\text{eff}} \frac{\overline{H}\,\overline{H}^*}{M}\right) \tag{28}$$

where

$$\rho_{\text{eff}} = \begin{cases} \frac{\rho T}{T_d - M}\left(\sqrt{\gamma}-\sqrt{\gamma-1}\right)^2, & \text{for } T_d > M \\ \frac{(\rho T)^2}{4M(M+\rho T)}, & \text{for } T_d = M \\ \frac{\rho T}{M-T_d}\left(\sqrt{-\gamma}-\sqrt{-\gamma+1}\right)^2, & \text{for } T_d < M \end{cases} \tag{29}$$

These formulas are especially revealing at high and low SNR. At high SNR, we have

$$\gamma = \frac{T_d}{T_d - M}$$

and at low SNR

$$\gamma = \frac{MT_d}{\rho T(T_d - M)}$$

so that we obtain the following results.

Corollary 1 (High and Low SNR):

1) At high SNR

$$\alpha = \frac{\sqrt{T_d}}{\sqrt{T_d}+\sqrt{M}}, \qquad \rho_{\text{eff}} = \frac{T}{\left(\sqrt{T_d}+\sqrt{M}\right)^2}\,\rho. \tag{30}$$

2) At low SNR

$$\alpha = \frac{1}{2}, \qquad \rho_{\text{eff}} = \frac{T^2}{4MT_d}\,\rho^2. \tag{31}$$

At low SNR, since $\alpha = 1/2$, *half* of the transmit energy $(\rho \cdot T)$ is devoted to training, and the effective SNR (and, consequently, the capacity) is quadratic in ρ.

C. Optimizing Over T_τ

All that remains is to determine the length of the training interval T_τ. We show that setting $T_\tau = M$ is optimal for any ρ and T (provided that we optimize ρ_τ and ρ_d). There is a simple intuitive explanation for this result. Increasing T_τ beyond M linearly decreases the bound through the $\frac{T-T_\tau}{T}$ term in (28), but only logarithmically increases the bound through the higher effective SNR ρ_{eff}. We, therefore, have a natural tendency to make T_τ as small as possible. Although making T_τ small loses accuracy in estimating H, we can compensate for this loss by increasing ρ_τ (even though this decreases ρ_d). We have the following result, which is the last step in our list of optimizations.

Theorem 3 (Optimal Training Interval): The optimal length of the training interval is $T_\tau = M$ for all ρ and T, and the capacity lower bound is

$$C_\tau \geqslant \mathrm{E}\,\frac{T-M}{T}\log\det\left(I_M + \rho_{\text{eff}}\frac{\overline{H}\,\overline{H}^*}{M}\right) \tag{32}$$

where

$$\rho_{\text{eff}} = \begin{cases} \frac{\rho T}{T-2M}\left(\sqrt{\gamma}-\sqrt{\gamma-1}\right)^2, & \text{for } T > 2M \\ \frac{\rho^2}{1+2\rho}, & \text{for } T = 2M \\ \frac{\rho T}{2M-T}\left(\sqrt{-\gamma}-\sqrt{-\gamma+1}\right)^2, & \text{for } T < 2M \end{cases}$$

$$\gamma = \frac{(M+\rho T)(T-M)}{\rho T(T-2M)}. \tag{33}$$

The optimal allocation of power is as given in (27) with $T_d = T - T_\tau = T - M$ and can be approximated at high SNR by

$$\alpha = \frac{\sqrt{T-M}}{\sqrt{T-M}+\sqrt{M}}, \quad \rho_{\text{eff}} = \frac{1}{\left(\sqrt{1-\frac{M}{T}}+\sqrt{\frac{M}{T}}\right)^2}\,\rho \tag{34}$$

and the power allocation becomes

$$\rho_d = \frac{\rho}{1-\frac{M}{T}+\sqrt{\left(1-\frac{M}{T}\right)\frac{M}{T}}}, \quad \rho_\tau = \frac{\rho}{\frac{M}{T}+\sqrt{\left(1-\frac{M}{T}\right)\frac{M}{T}}}. \tag{35}$$

To show this, we examine the case $T_d > M$ and omit the cases $T_d = M$ and $T_d < M$ since they are handled similarly. Let $Q = \min(M, N)$ and let λ denote an arbitrary nonzero eigenvalue of the matrix $\frac{\overline{H}\,\overline{H}^*}{M}$. Then we may rewrite (28) as

$$C_\tau \geqslant \underbrace{\frac{QT_d}{T}\,\mathrm{E}\log\left(1+\rho_{\text{eff}}\lambda\right)}_{C_t}$$

where the expectation is over λ. The behavior of C_t as a function of $T_d = T - T_\tau$ is studied. Differentiating C_t yields

$$\frac{dC_t}{dT_d} = \frac{Q}{T}\,\mathrm{E}\log\left(1+\rho_{\text{eff}}\lambda\right) + \frac{QT_d}{T}\,\frac{d\rho_{\text{eff}}}{dT_d}\,\mathrm{E}\left[\frac{\lambda}{1+\rho_{\text{eff}}\lambda}\right]. \tag{36}$$

After some algebraic manipulation of (25), it is readily verified that

$$\frac{d\rho_{\text{eff}}}{dT_d} = \frac{\rho T\left(\sqrt{\gamma}-\sqrt{\gamma-1}\right)^2}{(T_d-M)^2}\left(\frac{M\sqrt{\gamma}}{T_d\sqrt{\gamma-1}}-1\right)$$

which we plug into (36) and use the equality

$$1 - M\sqrt{\gamma}/(T_d\sqrt{\gamma-1}\,) = 1 - \sqrt{M(M+\rho T)/[T_d(\rho T + T_d)]}$$

to get

$$\frac{dC_t}{dT_d} = \frac{Q}{T}\,\mathrm{E}\Bigg[\log(1+\rho_{\text{eff}}\lambda) - \frac{\rho_{\text{eff}}\lambda}{1+\rho_{\text{eff}}\lambda}\,\frac{T_d}{T_d-M}\left(1-\sqrt{\frac{M(M+\rho T)}{T_d(\rho T+T_d)}}\right)\Bigg]. \tag{37}$$

The proof concludes by showing that $dC_t/dT_d > 0$; for then making T_d as large as possible (or, equivalently, T_τ as small as possible) maximizes C_t.

It suffices to show that the argument of the expectation in (37) is nonnegative for all $\lambda \geqslant 0$. Observe that because $T_d > M$

$$\frac{T_d}{T_d-M}\left(1-\sqrt{\frac{M(M+\rho T)}{T_d(\rho T+T_d)}}\right) < 1.$$

This is readily seen by isolating the term

$$\sqrt{M(M+\rho T)/[T_d(\rho T+T_d)]}$$

on the left-hnad side of the inequality and squaring both sides. From (37), it therefore suffices to show that

$$\log(1+\rho_{\text{eff}}\lambda) - \frac{\rho_{\text{eff}}\lambda}{1+\rho_{\text{eff}}\lambda} \geqslant 0, \qquad \lambda \geqslant 0.$$

But the function $\log(1+x) - x/(1+x) \geqslant 0$ because it is zero at $x = 0$ and its derivative is $x/(1+x)^2 \geqslant 0$ for all $x \geqslant 0$.

The formulas in (34) and (35) are verified by setting $T_d = T - M$ in (30). This concludes the proof.

This theorem shows that the optimal amount of training is the minimum possible $T_\tau = M$, provided that we allow the training and data powers to vary. In Section III-D, it is shown that if the constraint $\rho_\tau = \rho_d = \rho$ is imposed, the optimal amount of training may be greater than M.

We can also make some conclusions about the transmit powers.

Corollary 2 (Transmit Powers): The training and data power inequalities

$$\begin{aligned} &\rho_d < \rho < \rho_\tau \qquad (T > 2M) \\ &\rho_\tau < \rho < \rho_d \qquad (T < 2M) \\ &\rho_d = \rho = \rho_\tau \qquad (T = 2M) \end{aligned}$$

hold for all SNR ρ.

To show this, we concentrate on the case $T > 2M$, and omit the remaining two cases since they are similar. From the definition of α (23), we have

$$\rho_d = \frac{\alpha \rho T}{T - M}.$$

We need to show that $\rho_d < \rho$ or, equivalently

$$\frac{\alpha T}{T - M} < 1.$$

Using (27), we can transform this inequality into

$$\gamma - \sqrt{\gamma(\gamma - 1)} < \frac{T - M}{T}$$

or

$$\sqrt{\gamma(\gamma - 1)} > \gamma - \frac{T - M}{T}.$$

But this is readily verified by squaring both sides, cancelling common terms, and applying the formula for γ (33). We also need to show that $\rho_\tau > \rho$. We could again use (23) and show that

$$\frac{(1-\alpha)T}{M} > 1.$$

But it is simpler to argue that conservation of energy $\rho T = \rho_d T_d + \rho_\tau T_\tau$ where $T = T_d + T_\tau$ immediately implies that if $\rho_d < \rho$ then $\rho_\tau > \rho$, and conversely.

Thus, we spend more power for training when $T > 2M$, more power for data transmission when $T < 2M$, and the same power when $T = 2M$. We note that there have been some proposals for multiple-antenna differential modulation [14], [15] that use M transmit antennas and an effective block size of $T = 2M$. These proposals can be thought of as a natural extension of standard single-antenna differential phase-shift keying (DPSK), where the first half of the transmission (comprising M time samples across M transmit antennas) acts as a reference for the second half (also comprising M time samples). A differential scheme using orthogonal designs is proposed in [16]. In these proposals, both halves of the transmission are given equal power. But because $T = 2M$, Corollary 2 says that giving each half equal power maximizes the capacity lower bound. Thus, these differential proposals fortuitously follow the information-theoretic prescription that we derive here.

1) Low SNR: We know from Theorem 3 that the optimum training interval is $T_\tau = M$. Nevertheless, we show that at low SNR, the bound is actually not sensitive to the length of the training interval. We use Theorem 2, (28) and (29), and approximate

$$\left(\sqrt{\gamma} - \sqrt{\gamma - 1}\right)^2 \approx \frac{\rho T (T_d - M)}{4 M T_d}$$

for small ρ to obtain

$$\begin{aligned} C_\tau &\geqslant \frac{T_d}{T} \operatorname{E} \operatorname{tr} \log \left(I_M + \frac{T^2}{4 M T_d} \rho^2 \frac{\overline{H}\,\overline{H}^*}{M} \right) && (38) \\ &\approx \frac{T_d}{T} (\log e) \operatorname{E} \operatorname{tr} \left(\frac{T^2}{4 M T_d} \rho^2 \frac{\overline{H}\,\overline{H}^*}{M} \right) \\ &\approx \frac{T_d}{T} \frac{T^2 \log e}{4 M T_d} \rho^2 N \\ &= \frac{N T \log e}{4M} \rho^2 && (39) \end{aligned}$$

where in the first step we use $\log\det(\cdot) = \operatorname{tr}\log(\cdot)$, and in the second step we use the expansion $\log(I + A) = (\log e)(A - A^2/2 + A^3/3 - \cdots)$ for any matrix A with eigenvalues strictly inside the unit circle. Observe that the last expression is independent of T_τ. From Corollary 1, at low SNR optimum throughput occurs at $\alpha = \frac{1}{2}$. We, therefore, have the freedom to choose T_τ and ρ_τ in any way such that $\rho_d T_d = \rho_\tau T_\tau = \frac{1}{2}\rho T$. In particular, we may choose $\rho_\tau = \rho_d = \rho$ and $T_\tau = T_d = T/2$, which implies that when we choose equal training and data powers, half of the coherence interval should be spent training. The next section has more to say about optimizing T_τ when the training and data powers are equal.

At low power, the capacity lower bound (39) decays as ρ^2 because the effective SNR ρ_{eff} (31) decays as ρ^2; the quality of the channel estimate is very poor. The true channel capacity, however, (which does not necessarily require training to achieve) decays as ρ, rather than as ρ^2 [17], [18]. These simple power considerations therefore suggest that training and using the channel estimate as if it were correct is highly suboptimal when ρ is small.

D. Equal Training and Data Power

A communication system often does not have the luxury of varying the power during the training and data phases. If we assume that the training and data symbols are transmitted at the same power $\rho_\tau = \rho_d = \rho$ then (21) and (22) become

$$C_\tau \geqslant \operatorname{E} \frac{T - T_\tau}{T} \log\det \left(I_M + \frac{\rho^2 T_\tau / M}{1 + (1 + T_\tau/M)\rho} \frac{\overline{H}\,\overline{H}^*}{M} \right). \tag{40}$$

The effects and tradeoffs involving the training interval length T_τ can be inferred from the above formula. As we increase T_τ, our estimate of the channel improves and so

$$\rho_{\text{eff}} = \frac{\rho^2 T_\tau / M}{1 + (1 + T_\tau/M)\rho}$$

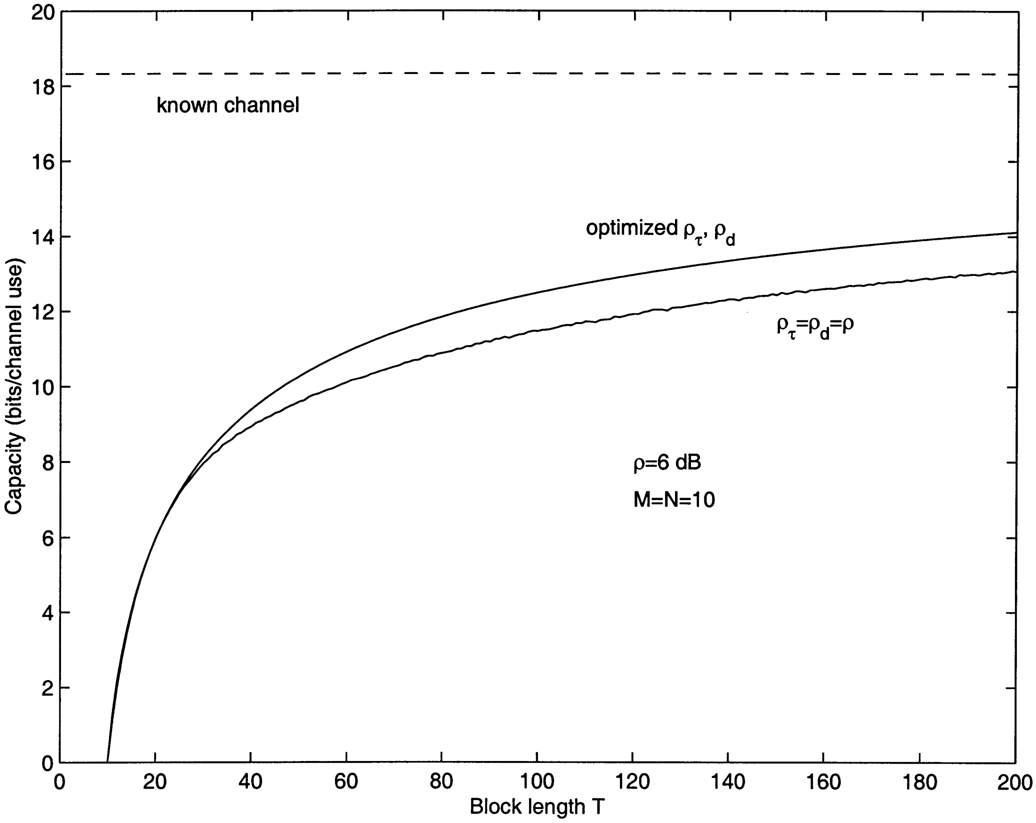

Fig. 1. The training-based lower bound on capacity as a function of T when SNR $\rho = 6$ dB and $M = N = 10$, for optimized ρ_τ and ρ_d (upper solid curve, (32)) and for $\rho_\tau = \rho$ (lower solid curve, (40) optimized for T_τ). The dashed line is the capacity when the receiver knows the channel.

increases, thereby increasing the capacity. On the other hand, as we increase T_τ the time available to transmit data decreases, thereby decreasing the capacity. Since the decrease in capacity is linear (through the coefficient $\frac{T-T_\tau}{T}$), whereas the increase in capacity is logarithmic (through ρ_{eff}), it follows that the length of the data transmission phase is a more precious resource than the effective SNR. Therefore, one may expect that it is possible to tolerate lower ρ_{eff} as long as T_d is long enough. Of course, the optimal value of T_τ in (40) depends on ρ, T, M, and N, and can be obtained by evaluating the lower bound in (40) (either analytically, see, e.g., [1], or via Monte Carlo simulation) for various values of T_τ. In fact, it can be shown that if the SNR is sufficiently high then $T_\tau = M$, and if the SNR is sufficiently low then $T_\tau = T/2$. In general, decreasing ρ requires increasing T_τ.

Some further insight into the tradeoff can be obtained by examining (40) at high and low SNRs.

1) At high SNR:

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\,\mathrm{logdet}\left(I_M+\frac{\rho}{1+\frac{M}{T_\tau}}\,\frac{\overline{H}\,\overline{H}^*}{M}\right). \tag{41}$$

Computing the optimal value of T_τ requires evaluating the expectation in the above inequality for $T_\tau = M, \ldots, T-1$.

2) At low SNR:

$$\begin{aligned} C_\tau &\geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\,\mathrm{tr}\log\left(I_M+\frac{\rho^2 T_\tau}{M}\,\frac{\overline{H}\,\overline{H}^*}{M}\right) \\ &\approx \frac{T-T_\tau}{T}\,\mathrm{E}\,\mathrm{tr}\,\frac{\rho^2 T_\tau \log e}{M}\cdot\frac{\overline{H}\,\overline{H}^*}{M} \\ &= \frac{NT_\tau(T-T_\tau)\log e}{MT}\,\rho^2. \end{aligned} \tag{42}$$

This expression is maximized by choosing $T_\tau = T/2$, from which we obtain

$$C_\tau \geqslant \frac{NT\log e}{4M}\,\rho^2. \tag{43}$$

This expression coincides with the expression obtained in Section III-C1. In other words, at low SNR, if we transmit the same power during training and data transmission, we need to devote half of the coherence interval to training, and the capacity is quadratic in ρ.

IV. Plots of Training Intervals and Capacities

Figs. 1 and 2 display the lower bound obtained as a function of the block length T for $M = N = 10$ when ρ_τ and ρ_d are optimized versus $\rho_\tau = \rho_d = \rho$. These figures assume that H has independent $\mathcal{CN}(0,\,1)$ entries. We see that approximately 5–10% gains in capacity are possible by allowing the training and data transmitted powers to vary. We also note that even when $T = 200$, we are approximately 15–20% from the capacity achieved when the receiver knows the channel. The curves for optimal ρ_τ and ρ_d were obtained by plotting (32) in Theorem 3, and the curves for $\rho_\tau = \rho_d = \rho$ were obtained by maximizing (40) over T_τ.

We know that if ρ_τ and ρ_d are optimized, then the optimal training interval $T_\tau = M$, but when the constraint $\rho_\tau = \rho_d = \rho$

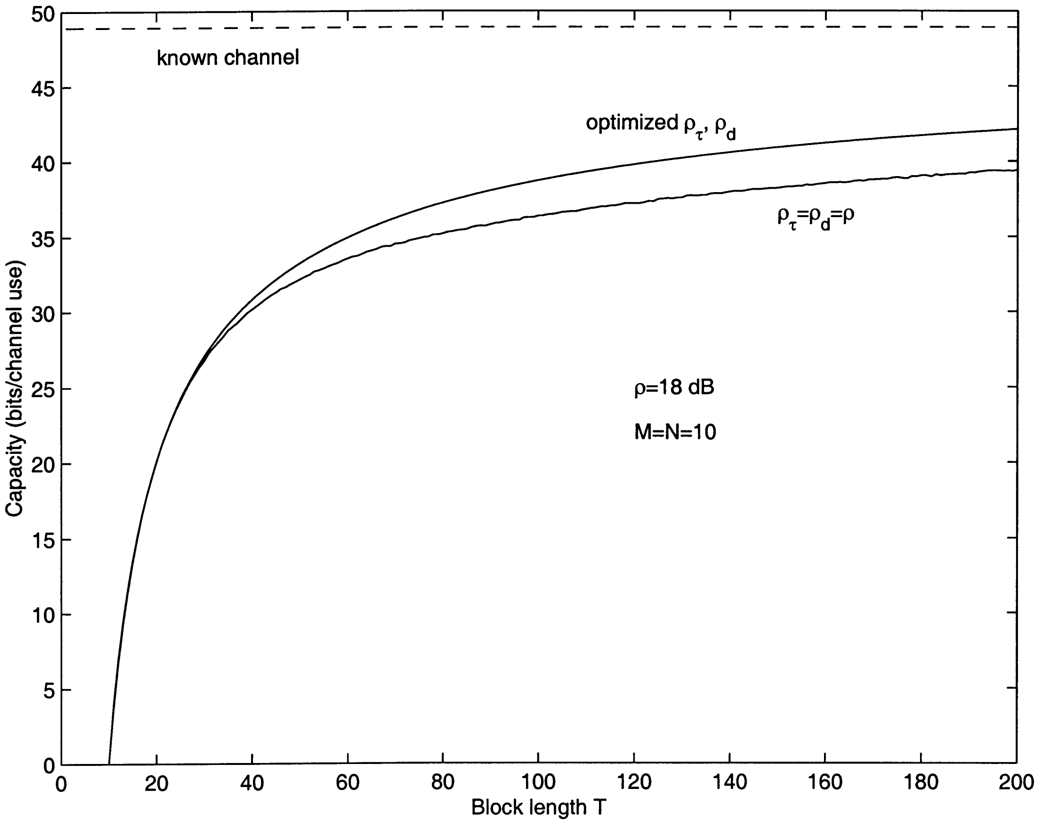

Fig. 2. Same as Fig. 1, except with $\rho = 18$ dB.

is imposed then $T_\tau \geqslant M$. Fig. 3 displays the T_τ that maximizes (40) for different values of ρ with $M = N = 10$. We see the trend that as the SNR decreases, the amount of training increases. It is shown in Section III-D that as $\rho \to 0$ the training increases until it reaches $T/2$.

Fig. 4 shows the variation of ρ_τ and ρ_d with the block length T for $\rho = 18$ dB and $M = N = 10$. We see the effects described in Corollary 2 where $\rho_\tau < \rho < \rho_d$ when $T < 2M = 20$ and $\rho_\tau = \rho_d = \rho$ when $T = 2M$ and $\rho_\tau > \rho > \rho_d$ when $T > 2M$. For sufficiently long T, the difference in SNR can apparently be more than 6 dB.

For a given SNR ρ, coherence interval T, and number of receive antennas N, we can calculate the capacity lower bound as a function of M. For $M \approx 1$, the training-based capacity is small because there are few antennas, and for $M \approx T$, the capacity is again small because we spend the entire coherence interval training. We can seek the value of M that maximizes this capacity. Figs. 5 and 6 show the capacity as a function of M for $\rho = 18$ dB, $N = 12$, and two different values of T. We see that the capacity when $T = 100$ peaks at $M \approx 15$ whereas it peaks at $M \approx 7$ when $T = 20$. We have included both optimized ρ_τ and ρ_d and equal $\rho_\tau = \rho_d = \rho$ for comparison. It is perhaps surprising that the number of transmit antennas that maximizes capacity often appears to be quite small. We see that choosing to train with the wrong number of antennas can severely hurt the data rate. This is especially true when $M \approx T$, where the capacity for the known channel is greatest, but the capacity for the system that trains all M antennas is least.

V. Discussion and Conclusion

The lower bounds on the capacity of multiple-antenna training-based schemes show that optimizing over the power allocation ρ_τ and ρ_d makes the optimum length of the training interval T_τ equal to M for all ρ and T. At high SNR, the resulting capacity lower bound is

$$C(\rho, T, M, N) \geqslant \left(1 - \frac{M}{T}\right) \mathrm{E} \log \cdot \det \left(I_M + \frac{1}{\left(\sqrt{1 - \frac{M}{T}} + \sqrt{\frac{M}{T}}\right)^2} \rho \frac{\overline{H}\,\overline{H}^*}{M} \right) \tag{44}$$

where $\overline{H}$ has independent $\mathcal{CN}(0, 1)$ entries.

If we require the power allocation for training and transmission to be the same, then the length of the training interval can be longer than M, although simulations at high SNR suggest that it is not much longer. As the SNR decreases, however, the training interval increases until at low SNR it converges to half the coherence interval.

The lower bounds on the capacity suggest that training-based schemes perform poorly when T is "close" to M. In fact, when $T = M$, the capacity bound is zero since the training phase occupies the entire coherence interval. Figs. 5 and 6 suggest that it is beneficial to use a training-based scheme with a smaller number of antennas $M' < M$. We may ask what is the optimal value of M' ? To answer this, we suppose that M antennas are

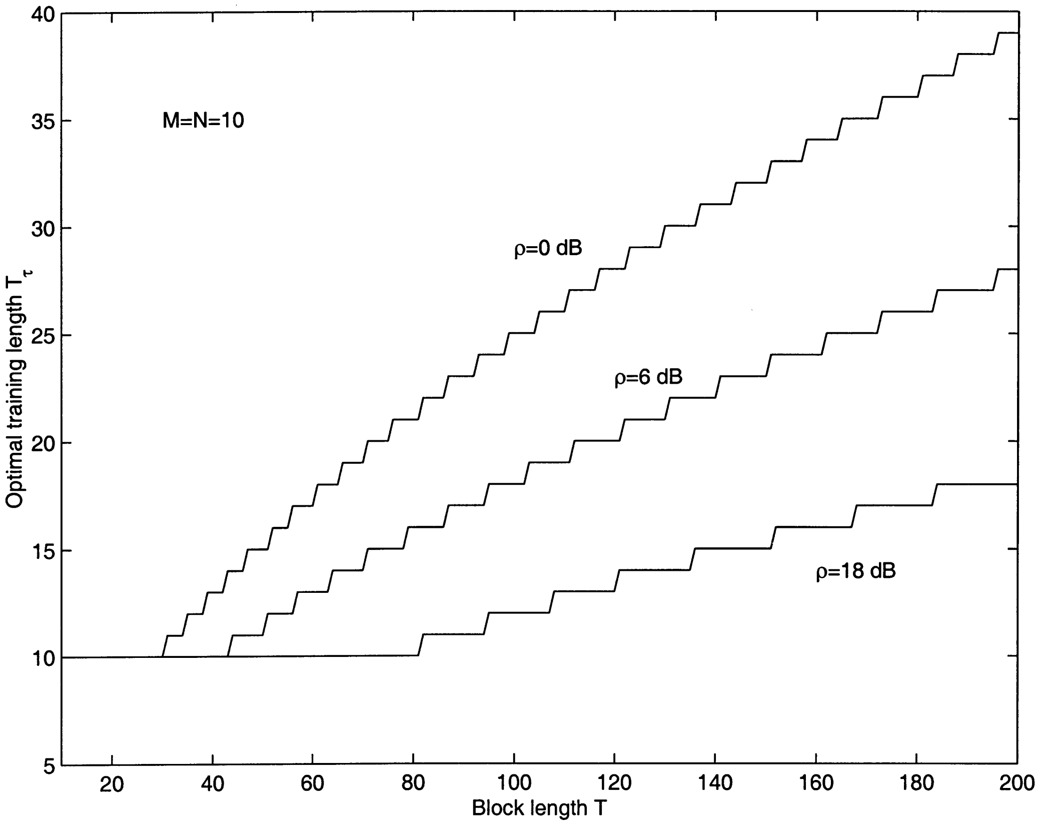

Fig. 3. The optimal amount of training T_τ as a function of block length T for three different SNRs ρ, for $M = N = 10$ and constraining the training and data powers to be equal $\rho_\tau = \rho_d = \rho$. The curves were made by numerically finding the T_τ that maximized (40).

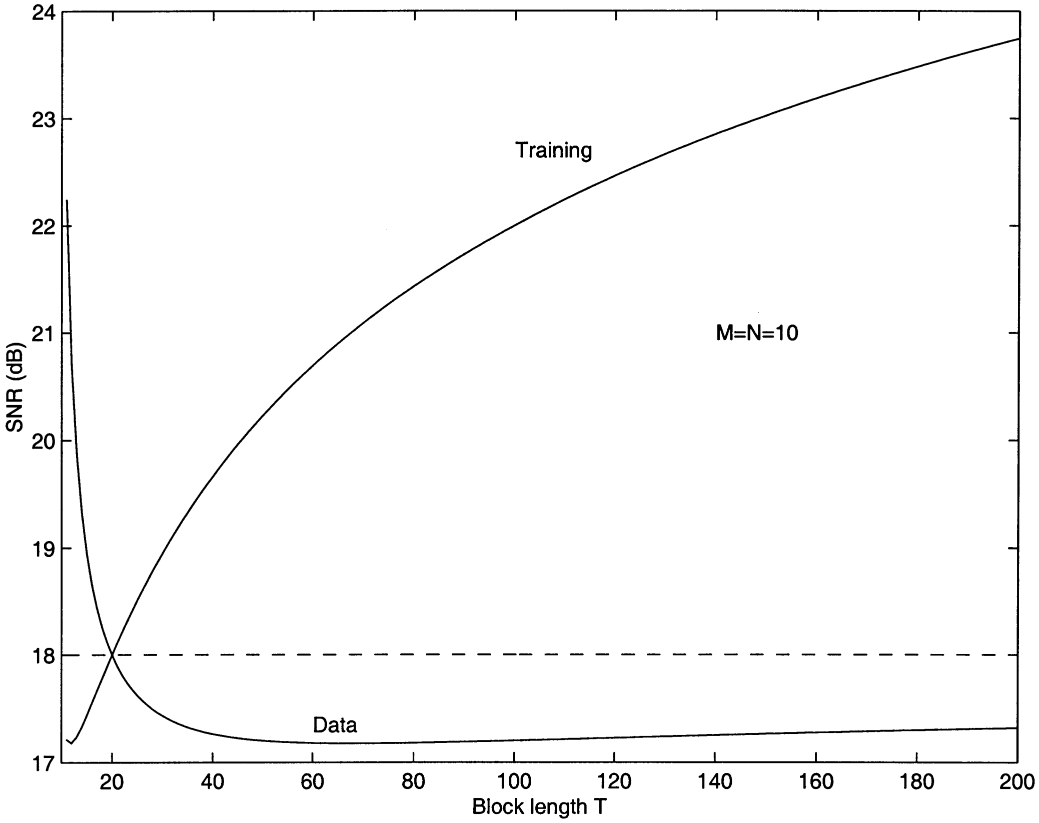

Fig. 4. The optimal power allocation ρ_τ (training) and ρ_d (data transmission) as a function of block length T for $\rho = 18$ dB (shown in the dashed line) with $M = N = 10$. These curves are drawn from Theorem 2 and (27) for $T_\tau = M$.

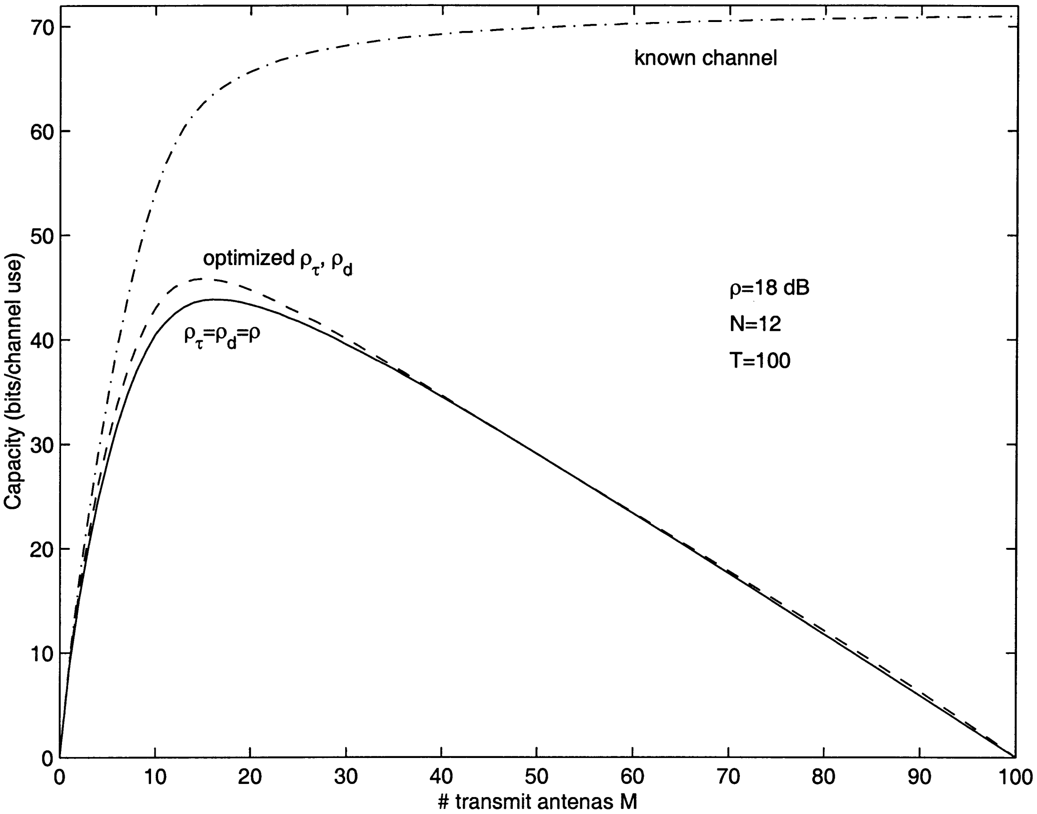

Fig. 5. Capacity as a function of number of transmit antennas M with $\rho = 18$ dB and $N = 12$ receive antennas. The solid line is optimized over T_τ for $\rho_\tau = \rho_d = \rho$ (see (40)), and the dashed line is optimized over the power allocation with $T_\tau = M$ (Theorem 3). The dash-dotted line is the capacity when the receiver knows the channel perfectly. The maximum throughput is attained at $M \approx 15$.

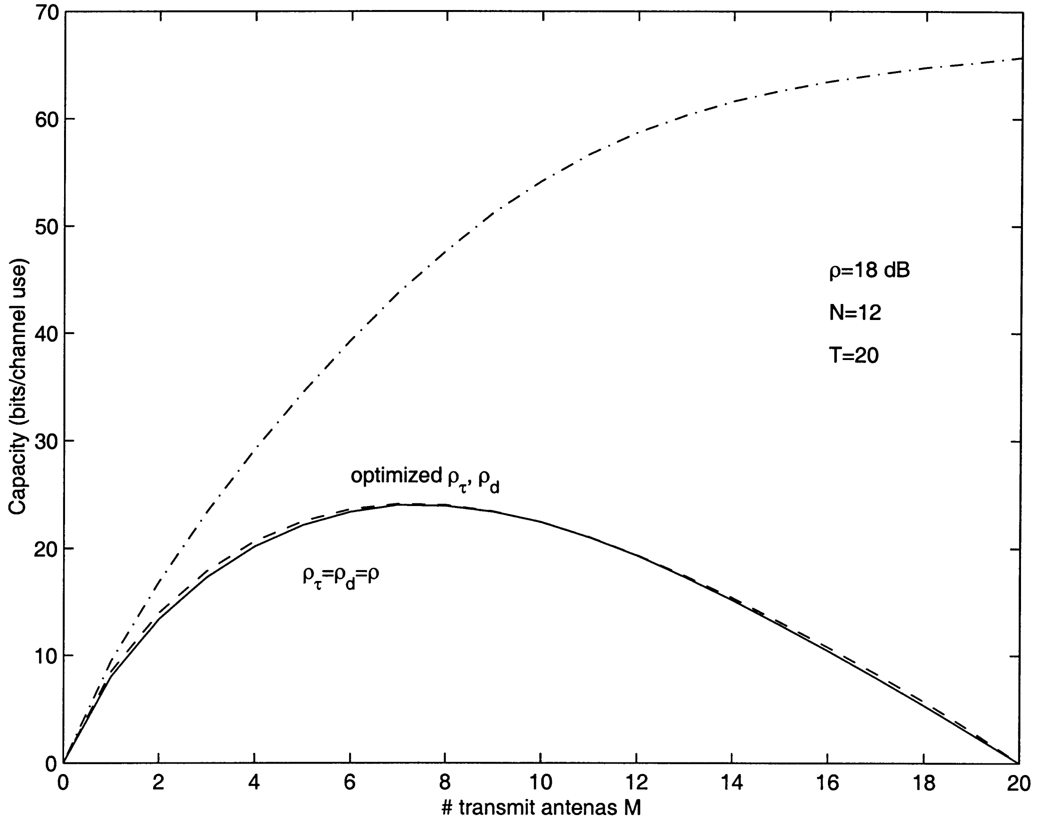

Fig. 6. Same as Fig. 5, except with $T = 20$. The maximum throughput is attained at $M \approx 7$. Observe that the difference between optimizing over ρ_τ and ρ_d versus setting $\rho_\tau = \rho_d = \rho$ is negligible.

available but we elect to use only $M' \leqslant M$ of them in a training-based scheme. Equation (44) is then rewritten as

$$C(\rho, T, M, N) \geqslant \max_{M' \leqslant M} \left(1 - \frac{M'}{T}\right) \mathrm{E} \log \cdot \det \left(I'_M + \frac{1}{\left(\sqrt{1-\frac{M'}{T}} + \sqrt{\frac{M'}{T}}\,\right)^2} \rho \frac{\overline{H}\,\overline{H}^*}{M'} \right). \quad (45)$$

Defining $Q = \min(M', N)$ and λ to be an arbitrary nonzero eigenvalue of

$$\frac{1}{(\sqrt{1-\frac{M'}{T}} + \sqrt{\frac{M'}{T}}\,)^2} \frac{\overline{H}\,\overline{H}^*}{M'}$$

we write

$$C(\rho, T, M, N) \geqslant \max_{M' \leqslant M} \left(1 - \frac{M'}{T}\right) Q\, \mathrm{E} \log(1 + \rho\lambda).$$

At high SNR, the leading term involving ρ becomes

$$C(\rho, T, M, N) \geqslant \max_{M' \leqslant M} \begin{cases} \left(1 - \frac{M'}{T}\right) M' \log \rho, & \text{if } M' \leqslant N \\ \left(1 - \frac{M'}{T}\right) N \log \rho, & \text{if } M' > N. \end{cases}$$

The expression $(1 - \frac{M'}{T}) M' \log \rho$ is maximized by the choice $M' = T/2$ when $\min(M, N) \geqslant T/2$, and by the choice $M' = \min(M, N)$ when $\min(M, N) < T/2$. This means that the expression is maximized when $M' = \min(M, N, T/2)$. The expression $(1 - \frac{M'}{T}) N \log \rho$, on the other hand, is maximized when $M' = N = \min(M, N)$ (since in this case $M > N$). Defining $K = \min(M, N, T/2)$, we conclude that

$$C(\rho, T, M, N) \geqslant \max \left[\left(1 - \frac{K}{T}\right) K \log \rho, \left(1 - \frac{\min(M, N)}{T}\right) \min(M, N) \log \rho \right].$$

When $\min(M, N) > T/2$ the first term is larger, and when $\min(M, N) \leqslant T/2$ the two terms are equal. Thus,

$$C(\rho, T, M, N) \geqslant \left(1 - \frac{K}{T}\right) K \log \rho. \quad (46)$$

This argument implies that, at high SNR, the optimal number of transmit antennas to use is $K = \min(M, N, T/2)$. We see indications of this result in Fig. 5 where the maximum throughput is attained at $M \approx 15$ versus the predicted high SNR value of $K = 12$, and in Fig. 6 at $M \approx 7$ versus the predicted $K = 10$.

We now ask whether the high-SNR bound (46) is tight? The answer to this question can be found in the recent work [19] of Zheng and Tse, where it is shown that at high SNR, the leading term of the actual channel capacity (*without* imposing any constraints such as training) is $\left(1 - \frac{K}{T}\right) K \log \rho$. Thus, in the leading SNR term (as $\rho \to \infty$), training-based schemes can be optimal, provided we use $K = \min(M, N, T/2)$ transmit antennas. (A similar conclusion is also drawn in [19].) Thus, it is possible to achieve capacity at high SNR by designing a transmitter/receiver pair that dedicates part of the transmission interval to training K antennas.

We note in Section III-C1 that, at low SNR, training and then using the channel estimate as if it were correct performs poorly because the effective SNR and capacity lower bound decay as ρ^2, whereas the actual capacity decays as ρ. The exact transition between what should be considered "high" SNR where this form of processing can yield acceptable performance versus "low" SNR where it does not, is not yet clear. Nevertheless, it is clear that a communication system that tries to achieve capacity at low SNR cannot rely on the accuracy of the channel estimate.

Appendix
Proof of Worst Case Noise Theorem

Consider the matrix-valued additive noise known channel

$$X = \sqrt{\frac{\rho}{M}} SH + V, \quad (A1)$$

where $H \in \mathcal{C}^{M \times N}$ is the known channel, $S \in \mathcal{C}^{1 \times M}$ is the transmitted signal, and $V \in \mathcal{C}^{1 \times N}$ is the additive noise. Assume further that the entries of S and V on the average have unit mean-square value, i.e.,

$$\mathrm{E} \frac{1}{M} SS^* = 1 \quad \text{and} \quad \mathrm{E} \frac{1}{N} VV^* = 1. \quad (A2)$$

The goal in this appendix is to find the worst case noise distribution for V in the sense that it minimizes the capacity of the channel (A1) subject to the power constraints (A2).

The arguments of [1], [2], which assume $R_V = I_N$, can be generalized in a straightforward manner to find the capacity of the channel (A1) when V has a zero-mean complex Gaussian distribution with variance $R_V = \mathrm{E}\, V^* V$ (additive Gaussian noise channel). The result is

$$C = \max_{R_S, \, \mathrm{tr}\, R_S = M} \mathrm{E} \log \det \left(I_N + \frac{\rho}{M} R_V^{-1} H^* R_S H \right). \quad (A3)$$

We obtain the worst case noise distribution when the noise V and the signal S are uncorrelated

$$\mathrm{E}\, S^* V = 0_{M \times N}. \quad (A4)$$

Let

$$C_{\text{worst}} = \inf_{p_V(\cdot), \, \mathrm{E}\, VV^* = N} \; \sup_{p_S(\cdot), \, \mathrm{E}\, SS^* = M} I(X; S | H).$$

Any particular distribution on V yields an upper bound on the worst case; choosing V to be zero-mean complex Gaussian with some covariance R_V and using (A3) yields

$$C_{\text{worst}} \leqslant \min_{R_V, \, \mathrm{tr}\, R_V = N} \; \max_{R_S, \, \mathrm{tr}\, R_S = M} \mathrm{E} \log \cdot \det \left(I_N + \frac{\rho}{M} R_V^{-1} H^* R_S H \right). \quad (A5)$$

To obtain a lower bound on C_{worst}, we compute the mutual information for the channel (A1) assuming that S is zero-mean complex Gaussian with covariance matrix R_S, but that the distribution on V is arbitrary. Thus,

$$\begin{aligned} I(X; S|H) &= h(S|H) - h(S|X, H) \\ &= \log \det \pi e R_S - h(S|X, H). \end{aligned}$$

Computing the conditional entropy $h(S|X, H)$ requires an explicit distribution on V. However, if the covariance matrix

$$\mathrm{cov}(S|X, H) = \mathrm{E}_{|X, H} (S - \mathrm{E}_{|X, H} S)^* (S - \mathrm{E}_{|X, H} S)$$

of the random variable $S_{|X,\,H}$ is known, $h(S|X,\,H)$ has the upper bound

$$h(S|X,\,H) \leqslant \mathrm{E}\log\det\pi e\,\mathrm{cov}(S|X,\,H)$$

since, among all random vectors with the same covariance matrix, the one with a Gaussian distribution has the largest entropy.

The following lemma gives a crucial property of $\mathrm{cov}(S|X,H)$. Its proof can be found in, for example, [20].

Lemma 1 (Minimum Covariance Property of $\mathrm{E}_{|X,\,H}S$*):* Let $\hat{S} = f(X,\,H)$ be *any* estimate of S given X and H. Then we have

$$\begin{aligned}\mathrm{cov}(S|X,\,H) &= \mathrm{E}(S-\mathrm{E}_{|X,\,H}S)^*(S-\mathrm{E}_{|X,\,H}S)\\ &\leqslant \mathrm{E}(S-\hat{S})^*(S-\hat{S}).\end{aligned} \tag{A6}$$

where the matrix inequality $A \leqslant B$ means that $B-A$ is positive semidefinite.

Substituting the LMMSE estimate $\hat{S} = XR_X^{-1}R_{XS}$ in this lemma yields

$$\begin{aligned}\mathrm{cov}(S|X,\,H) &\leqslant \mathrm{E}(S-XR_X^{-1}R_{XS})^*(S-XR_X^{-1}R_{XS})\\ &= R_S - R_{SX}R_X^{-1}R_{XS}.\end{aligned}$$

With the channel model (A1)–(A4), we see that

$$\begin{aligned}&R_S - R_{SX}R_X^{-1}R_{XS}\\ &\quad= R_S - \sqrt{\frac{\rho}{M}}\,R_SH\left(R_V+\frac{\rho}{M}\,H^*R_SH\right)^{-1}H^*R_S\sqrt{\frac{\rho}{M}}\\ &\quad= \left(R_S^{-1}+\frac{\rho}{M}\,HR_V^{-1}H^*\right)^{-1}.\end{aligned}$$

Thus,

$$\begin{aligned}h(S|X,\,H) &\leqslant \mathrm{E}\log\det\pi e\left(R_S^{-1}+\frac{\rho}{M}\,HR_V^{-1}H^*\right)^{-1}\\ &= \mathrm{E}\log\det\pi eR_S\left(I_N+\frac{\rho}{M}\,R_V^{-1}H^*R_SH\right)^{-1}\end{aligned}$$

from which it follows that, when S is complex Gaussian distributed, then for any distribution on V we have

$$I(X;S|H) \geqslant \mathrm{E}\log\det\left(I_N+\frac{\rho}{M}\,R_V^{-1}H^*R_SH\right). \tag{A7}$$

Since the above inequality holds for any R_S and R_V, we therefore have

$$\begin{aligned}C_{\mathrm{worst}} \geqslant &\min_{R_V,\,\mathrm{tr}\,R_V=N}\ \max_{R_S,\,\mathrm{tr}\,R_S=M}\mathrm{E}\log\\ &\cdot\det\left(I_N+\frac{\rho}{M}\,R_V^{-1}H^*R_SH\right).\end{aligned} \tag{A8}$$

The combination of this inequality and (A5) yields

$$\begin{aligned}C_{\mathrm{worst}} = &\min_{R_V,\,\mathrm{tr}\,R_V=N}\ \max_{R_S,\,\mathrm{tr}\,R_S=M}\mathrm{E}\log\\ &\cdot\det\left(I_N+\frac{\rho}{M}\,R_V^{-1}H^*R_SH\right).\end{aligned} \tag{A9}$$

To prove the inequalities in (12), we note that the inequality on the left follows from the fact that in an additive Gaussian noise channel the mutual-information-maximizing distribution on S is Gaussian. The inequality on the right follows from (A7), where S is Gaussian.

All that remains to be done is to compute the optimizing $R_{V,\,\mathrm{opt}}$ and $R_{S,\,\mathrm{opt}}$, when H is rotationally invariant. Consider first $R_{S,\,\mathrm{opt}}$. There is no loss of generality in assuming that R_S is diagonal: if not, take its eigenvalue decomposition $R_S = U\Lambda_sU^*$, where U is unitary and Λ_s is diagonal, and note that U^*H has the same distribution as H because H is left rotationally invariant. Now suppose that $R_{S,\,\mathrm{opt}}$ is diagonal with possibly unequal entries. Then form a new covariance matrix

$$R_S = \frac{1}{M!}\sum_{m=1}^{M!}P_mR_{S,\,\mathrm{opt}}P_m^* = I_M$$

where the $P_1,\,\ldots,\,P_{M!}$ are all possible $M\times M$ permutation matrices. Since the "expected log-det" function in (A9) is concave in R_S (see also [1]), the value of the function cannot decrease with the new covariance. We, therefore, conclude that $R_{S,\,\mathrm{opt}} = I_M$. A similar argument holds for $R_{V,\,\mathrm{opt}}$ because the "expected log-det" function in (A9) is convex in R_V.

References

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Europ. Trans. Telecomm.*, vol. 10, pp. 585–595, Nov. 1999.

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

[3] T. L. Marzetta, "BLAST training: Estimating channel characteristics for high-capacity space-time wireless," in *Proc. 37th Annu. Allerton Conf. Communications, Control, and Computing*, Sept. 22–24, 1999.

[4] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electron. Lett.*, vol. 35, pp. 14-–16, Jan. 1999.

[5] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1841–1852, Nov. 1999.

[6] W. C. Jakes, *Microwave Mobile Communications*. Piscataway, NJ: IEEE Press, 1993.

[7] N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.

[8] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1520–1529, Sept. 1996.

[9] M. Medard, "The effect upon channel capacity in wireless communication of perfect and imperfect knowledge of the channel," *IEEE Trans. Inform. Theory*, vol. 46, pp. 933–946, May 2000.

[10] M. Pinsker, "Calculation of the rate of information production by means of stationary random processes and the capacity of a stationary channel," *Dokl. Akad. Nauk USSR*, vol. 111, pp. 753–756, Sept. 1956.

[11] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Inform. Contr.*, vol. 37, pp. 34–39, Sept. 1978.

[12] R. McEliece and W. Stark, "An information theoretic study of communication in the presence of jamming," in *Proc. Int. Conf. Communication*, 1981, pp. 45.3.1–45.3.5.

[13] J.-C. Guey, M. Fitz, M. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, pp. 527–537, Apr. 1999.

[14] B. Hochwald and W. Sweldens, "Differential unitary space time modulation," *IEEE Trans. Commun.*, vol. 48, pp. 2041–2052, Dec. 2000. [Online]. Available: http://mars.bell-labs.com.

[15] B. Hughes, "Differential space-time modulation," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2567–2578, Nov. 2000.

[16] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity," *J. Sel. Area Comm.*, vol. 46, pp. 1169–1174, July 2000.

[17] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2619–2692, Oct. 1999.

[18] I. C. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1290–1301, May 2001.

[19] L. Zheng and D. Tse, "Communication on the Grassman manifold: A geometric approach to the noncoherent multiantenna channel," *IEEE Trans. Inform. Theory*, vol. 48, pp. 359–383, Feb. 2002.

[20] T. Söderström and P. Stoica, *System Identification*. London, U.K.: Prentice-Hall, 1989.

[21] S. N. Crozier, D. D. Falconer, and S. A. Mahmoud, "Least sum of squared errors (LSSE) channel estimation," *Proc Inst. Elec. Eng. Pt. F: Radar and Signal Processing*, vol. 138, pp. 371–378, Aug. 1991.