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

BABAK HASSIBI BERTRAND M. HOCHWALD
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974
{hassibi,hochwald}@bell-labs.com

August 30, 2000

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 use an information-theoretic approach to compute the optimal amount of training as a function of the received signal-to-noise ratio, 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. As side results, we obtain the worst-case power-constrained additive noise in a matrix-valued additive noise channel, and show that training-based schemes are highly suboptimal at low SNR.

Index terms—BLAST, space-time coding, transmit diversity, receive diversity, high-rate wireless communications

1 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 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 multi-antenna wireless link with M transmit antennas, N receive antennas, coherence interval of length T (in

symbols), and 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 [2] where an experimental prototype has achieved 20 bits/sec/Hz data rates with 8 transmit and 12 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 at low SNR or when the coherence interval T is only slightly larger than the number of transmit antennas M. 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 4. 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.

2 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 which it holds for another interval T, and so on. This is an appropriate model for TDMA- or frequency-hopping-based systems, and is a tractable approximation of a continuously fading channel model such as Jakes' [4]. 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 antennas, and V is

a $T \times N$ matrix of additive noise. The matrices H and V both comprise independent random variables with whose mean-square is unity. We also assume that the entries of the transmitted signal S have unit mean-square. 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.

2.1 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 learn H. In particular, training-based schemes are composed of the following two phases.

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

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

   where S_τ is the matrix of training symbols sent over T_τ time samples and known to the receiver, and ρ_τ is the SNR during the training phase. (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). The observed signal matrix $X_\tau \in \mathcal{C}^{T_\tau \times N}$ and S_τ are used to construct an estimate of the channel

$$\hat{H} = f(X_\tau, S_\tau). \tag{3}$$

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

$$\hat{H} = \sqrt{\frac{M}{\rho_\tau}} \left(S_\tau^* S_\tau\right)^{-1} S_\tau^* X_\tau, \qquad \hat{H} = \sqrt{\frac{M}{\rho_\tau}} \left(\frac{M}{\rho_\tau} I_M + S_\tau^* S_\tau\right)^{-1} S_\tau^* X_\tau. \tag{4}$$

   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 $T_\tau \geqslant M$.

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

$$X_d = \sqrt{\frac{\rho_d}{M}} S_d H + V_d, \quad S_d \in \mathcal{C}^{T_d \times M}, \quad \mathrm{E}\,\mathrm{tr}\, S_d S_d^* = M T_d \tag{5}$$

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 (5) has an expectation. The estimate of the channel $\hat{H}$ is used to recover S_d. This is written formally as

$$X_d = \sqrt{\frac{\rho_d}{M}} S_d \hat{H} + \underbrace{\sqrt{\frac{\rho_d}{M}} S_d \tilde{H} + V_d}_{V_d'}, \tag{6}$$

where $\tilde{H} = H - \hat{H}$ is the channel estimation error.

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

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

Conservation of time and energy yield

$$T = T_\tau + T_d, \qquad \rho T = \rho_\tau T_\tau + \rho_d T_d. \tag{7}$$

Within the data transmission interval the estimate $\hat{H}$ is used to recover the data. It is clear that increasing T_τ improves the estimate $\hat{H}$, 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. In this note, we compute T_τ to optimize the tradeoff of accuracy of $\hat{H}$ versus the length of the data transmission interval T_d.

3 Capacity and Capacity Bounds

In any training-based scheme, the capacity in bits/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

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

Now

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

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

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

Strictly speaking, as long the estimate of the channel matrix $\hat{H} = f(X_\tau, S_\tau)$ does not "throw away" information, the choice of the channel estimate in (6) does not affect the capacity because the capacity depends only on the conditional distribution of H given S_τ and X_τ. But most practical data transmission schemes that employ training do throw away information because they use the estimate $\hat{H}$ as if it were correct. We assume that such a scheme is employed.

In particular, we find a lower bound on the capacity by choosing a particular estimate of the channel. 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. We may write

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

where $\tilde{H} = H - \hat{H}$ is the zero-mean estimation error. By well-known properties of the conditional mean, $\hat{H}$ and $\tilde{H}$ are uncorrelated.

From (6), during the data transmission phase we may write

$$X_d = \sqrt{\frac{\rho_d}{M}} S_d \hat{H} + V_d' \tag{10}$$

where V_d' combines the additive noise and residual channel estimation error. The estimate $\hat{H} = f(X_\tau, S_\tau)$ is known and assumed by the training-based scheme to be correct; hence, the channel capacity of a training-based scheme is the same as the capacity of a *known channel* system, subject to additive noise with the power

constraint

$$\sigma_{V'}^2 = \frac{1}{NT_d}\mathrm{tr}\,\mathrm{E}\,V_d^{'}V_d^{'*} = \frac{1}{NT_d}\mathrm{E}\,\mathrm{tr}\,\left[\frac{\rho_d}{M}\tilde{H}\tilde{H}^*S_d^*S_d\right] + \frac{1}{NT_d}\mathrm{E}\,\mathrm{tr}\,V_dV_d^* = \frac{\rho_d}{MNT_d}\mathrm{tr}\,\left[\mathrm{E}\,(\tilde{H}\tilde{H}^*)\mathrm{E}\,(S_d^*S_d)\right] + 1. \quad (11)$$

There are two important differences between (10) and (1). In (10) 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 (10) it is possibly neither. Finding the capacity of a training-based scheme requires us to examine the worst effect the additive noise can have during data transmission. We therefore wish to find

$$C_{\mathrm{worst}} = \inf_{p_{V_d^{'}}(\cdot),\mathrm{tr}\,\mathrm{E}\,V_d^{'}V_d^{'*}=NT_d} \sup_{p_{S_d}(\cdot),\mathrm{tr}\,\mathrm{E}\,S_dS_d^{'*}=MT_d} I(X_d;S_d|\hat{H}).$$

A similar argument for lower-bounding the mutual information in a scalar and multiple-access wireless channel is given in [5]. The worst-case noise is the content of the next theorem, which is proven in Appendix A.

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

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

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, $V \sim \mathcal{CN}(0, R_{V,\mathrm{opt}})$, where $R_{V,\mathrm{opt}}$ is the minimizing noise covariance in

$$C_{\mathrm{worst}} = \min_{R_V,\mathrm{tr}\,R_V=N} \max_{R_S,\mathrm{tr}\,R_S=M} \mathrm{E}\,\log\det\left(I_N + \frac{\rho}{M}R_V^{-1}H^*R_SH\right). \quad (12)$$

We also have the minimax property

$$I_{V\sim\mathcal{CN}(0,R_{V,\mathrm{opt}}),S}(X;S) \;\leqslant\; I_{V\sim\mathcal{CN}(0,R_{V,\mathrm{opt}}),S\sim\mathcal{CN}(0,R_{S,\mathrm{opt}})}(X;S) = C_{\mathrm{worst}} \;\leqslant\; I_{V,S\sim\mathcal{CN}(0,R_{S,\mathrm{opt}})}(X;S), \tag{13}$$

where $R_{S,\mathrm{opt}}$ is the maximizing signal covariance matrix in (12). 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

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

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

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

When the additive noise $V_d^{'}$ and signal S_d are uncorrelated Theorem 1 shows that the worst-case additive noise is zero-mean temporally white Gaussian noise of an appropriate covariance matrix $R_{V,\mathrm{opt}}$ with normalization $\mathrm{tr}\, R_{V,\mathrm{opt}} = N\sigma_{V'}^2$. Because $\mathrm{E}\, S_d^* S_d = T_d R_S$, equation (11) becomes

$$\begin{aligned} \sigma_{V'}^2 &= 1 + \frac{\rho_d}{MNT_d}\mathrm{tr}\,\left[(\mathrm{E}\,\tilde{H}\tilde{H}^*)T_d R_S\right] \\ &= 1 + \rho_d \sigma_{\tilde{H},R_S}^2, \end{aligned} \tag{14}$$

where $\sigma_{\tilde{H},R_S}^2 \triangleq \frac{1}{NM}\mathrm{E}\,\mathrm{tr}\,\tilde{H}^* R_S \tilde{H}$.

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

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

$$= \quad 0 \quad \text{since } \mathrm{E}_{\,|X_\tau,S_\tau}(H-\hat{H}) = 0.$$

The MMSE estimate is the only estimate with this property.

The noise term V_d' in (10), 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 temporally white additive Gaussian noise with the same power constraint $\mathrm{tr}\, R_{V,\mathrm{opt}} = N(1+\rho_d\sigma^2_{\tilde{H},R_S})$. Using (12), we may therefore write

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

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 $\sigma^2_{\hat{H}} = \frac{1}{NM}\mathrm{E}\,\mathrm{tr}\,\hat{H}^*\hat{H}$. By the orthogonality principle for MMSE estimates,

$$\sigma^2_{\hat{H}} = 1 - \sigma^2_{\tilde{H}}, \tag{15}$$

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

$$\bar{H} \triangleq \frac{1}{\sigma_{\hat{H}}}\hat{H}.$$

We may write the capacity bound as

$$C_\tau \geqslant \min_{R_V,\mathrm{tr}\, R_V=N} \max_{R_S,\mathrm{tr}\, R_S=M} \mathrm{E}\, \frac{T-T_\tau}{T} \log\det\left(I_N + \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H},R_S}} \frac{R_V^{-1}\bar{H}^* R_S \bar{H}}{M}\right). \tag{16}$$

The ratio

$$\rho_{\mathrm{eff}} = \frac{\rho_d\sigma^2_{\hat{H}}}{1+\rho_d\sigma^2_{\tilde{H},R_S}} \tag{17}$$

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:

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

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

where $R_{H X_\tau} = \mathrm{E}\,(\operatorname{vec} H)(\operatorname{vec} X_\tau)^*$ 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 (4).) 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 $\bar{H}$, 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 (16) depends on the distribution of $\bar{H}$ which, in turn, depends on the training signal 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. Since we are interested in training-based schemes, the antennas are to be used as if the channel were learned perfectly at the receiver; thus, we choose $R_S = I_M$ (see [1]). Theorem 1 says that $R_S = I_M$ is optimal when the distribution of $\hat{H}$ is left rotationally invariant. Section 3.1 shows that the choice of S_τ that maximizes ρ_{eff} gives $\hat{H}$ this property. With $R_S = I_M$, we have

$$C_\tau \geqslant \mathrm{E}\,\frac{T - T_\tau}{T} \log\det\left(I_N + \frac{\rho_d \sigma_{\hat{H}}^2}{1 + \rho_d \sigma_{\tilde{H}}^2} \frac{\bar{H}^* \bar{H}}{M}\right). \tag{18}$$

Finally, we note from Theorem 1 that the bounds (16) and (18) are tight if the MMSE estimate of H is used in the training phase, and V_d' in (6) is Gaussian. However, $V_d' = \sqrt{\rho_d / M}\, S_d \tilde{H} + V_d$ is not, in general, Gaussian. But because V_d is Gaussian, V_d' becomes Gaussian as $\rho_d \to 0$. Hence the bounds (16) and (18) become tight at low SNR ρ. In Section 3.3.1 we use this tightness to conclude that training is suboptimal at low SNR. In Section 5 we show that these bounds are also tight at high SNR. We therefore expect these bounds to be reasonably tight for a wide range of SNR's.

3.1 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 (18) 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}

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

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

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

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

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

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

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 $\lambda_1 = \ldots = \lambda_M = T_\tau$. This yields

$$S_\tau^* S_\tau = T_\tau I_M, \tag{19}$$

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.

With this choice of training signal, we obtain

$$\sigma^2_{\tilde{H}} = \frac{1}{1+\frac{\rho_\tau}{M}T_\tau} \quad \text{and} \quad \sigma^2_{\hat{H}} = \frac{\frac{\rho_\tau}{M}T_\tau}{1+\frac{\rho_\tau}{M}T_\tau}. \tag{20}$$

In fact, we have the stronger result

$$R_{\tilde{H}} = \frac{1}{1+\frac{\rho_\tau}{M}T_\tau} I_M \otimes I_N \quad \text{and} \quad R_{\hat{H}} = \frac{\frac{\rho_\tau}{M}T_\tau}{1+\frac{\rho_\tau}{M}T_\tau} I_M \otimes I_N \tag{21}$$

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

Thus, (18) can be written as

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M + \rho_{\text{eff}}\frac{\bar{H}\bar{H}^*}{M}\right), \tag{22}$$

where

$$\rho_{\text{eff}} = \frac{\rho_d\rho_\tau T_\tau}{M(1+\rho_d)+\rho_\tau T_\tau}, \tag{23}$$

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

3.2 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,

$$\rho_d T_d = \alpha\rho T, \quad \rho_\tau T_\tau = (1-\alpha)\rho T, \qquad 0<\alpha<1. \tag{24}$$

Therefore we may write

$$\rho_{\text{eff}} = \frac{\rho_d\rho_\tau T_\tau}{M(1+\rho_d)+\rho_\tau T_\tau} \quad = \quad \frac{\alpha\frac{\rho T}{T_d}\cdot(1-\alpha)\rho T}{M(1+\alpha\frac{\rho T}{T_d})+(1-\alpha)\rho T}$$

$$
\begin{aligned}
&= \frac{(\rho T)^2}{T_d} \cdot \frac{\alpha(1-\alpha)}{M + \rho T - \rho T(1-\frac{M}{T_d})\alpha} \\
&= \frac{\rho T}{T_d - M} \cdot \frac{\alpha(1-\alpha)}{-\alpha + \frac{M+\rho T}{\rho T(1-\frac{M}{T_d})}}.
\end{aligned}
$$

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

$$\alpha = \frac{1}{2}, \tag{25}$$

and therefore that

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

2. $T_d > M$: We write

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

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

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

3. $T_d < M$: We write

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

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

$$\rho_{\text{eff}} = \frac{\rho T}{M - T_d}(\sqrt{-\gamma} - \sqrt{-\gamma + 1})^2. \tag{27}$$

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*

$$\alpha = \begin{cases} \gamma - \sqrt{\gamma(\gamma - 1)} & \text{for } T_d > M \\ \frac{1}{2} & \text{for } T_d = M \\ \gamma + \sqrt{\gamma(\gamma - 1)} & \text{for } T_d < M \end{cases} \tag{28}$$

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

$$C_\tau \geqslant \mathrm{E}\, \frac{T - T_\tau}{T} \log\det\left(I_M + \rho_{\text{eff}} \frac{\bar{H}\bar{H}^*}{M}\right), \tag{29}$$

where

$$\rho_{\text{eff}} = \begin{cases} \frac{\rho T}{T_d - M}(\sqrt{\gamma} - \sqrt{\gamma - 1})^2 & \text{for } T_d > M \\ \frac{(\rho T)^2}{4M(M + \rho T)} & \text{for } T_d = M \\ \frac{\rho T}{M - T_d}(\sqrt{-\gamma} - \sqrt{-\gamma + 1})^2 & \text{for } T_d < M \end{cases} \tag{30}$$

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*

$$\alpha = \frac{\sqrt{T_d}}{\sqrt{T_d} + \sqrt{M}}, \qquad \rho_{\text{eff}} = \frac{T}{(\sqrt{T_d} + \sqrt{M})^2}\rho. \tag{31}$$

2. At low SNR

$$\alpha = \frac{1}{2}, \qquad \rho_{\text{eff}} = \frac{T^2}{4MT_d}\rho^2. \tag{32}$$

When $T_d = M$, we see that $\rho_{\text{eff}} = (T/4M)\rho$ at high SNR, whereas $\rho_{\text{eff}} = (T^2/4M^2)\rho^2$ at low SNR. 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.

3.3 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 capacity through the $\frac{T-T_\tau}{T}$ term in (29), but only logarithmically increases the capacity 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*

$$C_\tau \geqslant \mathrm{E}\,\frac{T-M}{T}\log\det\left(I_M + \rho_{\text{eff}}\frac{\bar{H}\bar{H}^*}{M}\right), \tag{33}$$

where

$$\rho_{\text{eff}} = \begin{cases} \frac{\rho T}{T-2M}(\sqrt{\gamma}-\sqrt{\gamma-1})^2 & \text{for } T > 2M \\ \frac{\rho^2}{1+2\rho} & \text{for } T = 2M \\ \frac{\rho T}{2M-T}(\sqrt{-\gamma}-\sqrt{-\gamma+1})^2 & \text{for } T < 2M \end{cases}, \qquad \gamma = \frac{(M+\rho T)(T-M)}{\rho T(T-2M)}. \tag{34}$$

The optimal allocation of power is as given in (28) with $T_d = T - T_\tau = T - M$ and can be approximated at high SNR by

$$\alpha = \frac{\sqrt{T-M}}{\sqrt{T-M}+\sqrt{M}}, \qquad \rho_{\text{eff}} = \frac{1}{(\sqrt{1-\frac{M}{T}}+\sqrt{\frac{M}{T}})^2}\rho \tag{35}$$

and the power allocation becomes

$$\rho_d = \frac{\rho}{1-\frac{M}{T}+\sqrt{(1-\frac{M}{T})\frac{M}{T}}}, \quad \rho_\tau = \frac{\rho}{\frac{M}{T}+\sqrt{(1-\frac{M}{T})\frac{M}{T}}}. \tag{36}$$

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 $\frac{\bar{H}\bar{H}^*}{M}$.

Then we may rewrite (29) 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

$$\frac{dC_t}{dT_d} = \frac{Q}{T}\mathrm{E}\,\log\left(1+\rho_{\text{eff}}\lambda\right) + \frac{QT_d}{T}\frac{d\rho_{\text{eff}}}{dT_d}\mathrm{E}\left[\frac{\lambda}{1+\rho_{\text{eff}}\lambda}\right]. \tag{37}$$

After some algebraic manipulation of (26), it is readily verified that

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

which we plug into (37) 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

$$\frac{dC_t}{dT_d} = \frac{Q}{T}\mathrm{E}\left[\log(1+\rho_{\text{eff}}\lambda) - \frac{\rho_{\text{eff}}\lambda}{1+\rho_{\text{eff}}\lambda}\frac{T_d}{T_d-M}\left(1-\sqrt{\frac{M(M+\rho T)}{T_d(\rho T+T_d)}}\right)\right]. \tag{38}$$

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 (38) 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 side of the inequality and squaring both sides. From (38), 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 (35) and (36) are verified by setting $T_d = T - M$ in (31). 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 3.4 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 (24), 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 (28), 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 (34). We also need to show that $\rho_\tau > \rho$. We could again use (24) 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 [6], [7] 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 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 [8]. 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 is *optimal* in the sense of maximizing the capacity lower bound. Thus, these differential proposals fortuitously follow the information-theoretic prescription that we derive here.

3.3.1 Low SNR

We know from Theorem 3 that the optimum training interval is $T_\tau = M$. Nevertheless, we show that at low SNR the capacity is actually not sensitive to the length of the training interval. We use Theorem 2, equations (29) and (30), and approximate

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

for small ρ to obtain

$$\begin{aligned} C_\tau &\geqslant \frac{T_d}{T} \mathrm{E}\,\mathrm{tr}\,\log\left(I_M + \frac{T^2}{4MT_d}\rho^2 \frac{\bar{H}\bar{H}^*}{M}\right) && (39) \\ &\approx \frac{T_d}{T}(\log e)\mathrm{E}\,\mathrm{tr}\left(\frac{T^2}{4MT_d}\rho^2 \frac{\bar{H}\bar{H}^*}{M}\right) \\ &\approx \frac{T_d}{T}\frac{T^2 \log e}{4MT_d}\rho^2 N \\ &= \frac{NT\log e}{4M}\rho^2, && (40) \end{aligned}$$

where in the first step we use $\log\det(\cdot) = \mathrm{tr}\,\log(\cdot)$, 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.

The paragraph before Section 3.1 argues that our capacity lower bound (39) should be tight at low SNR. We therefore infer that, at low power, the capacity with training is given by (40) and decays as ρ^2. However,

the true channel capacity (which does not necessarily require training to achieve) decays as ρ [9], [10]. We therefore must conclude that training is highly suboptimal when ρ is small.

3.4 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 (22) and (23) become

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M + \frac{\rho^2 T_\tau/M}{1+(1+T_\tau/M)\rho}\frac{\bar{H}\bar{H}^*}{M}\right). \tag{41}$$

The effects and trade-offs 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 $\rho_{\text{eff}} = \frac{\rho^2 T_\tau/M}{1+(1+T_\tau/M)\rho}$ 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 (41) depends on ρ, T, M, and N, and can be obtained by evaluating the lower bound in (41) (either analytically, see, e.g., [1], or via Monte Carlo simulation) for various values of T_τ.

Some further insight into the trade-off can be obtained by examining (41) at high and low SNR's.

1. At high SNR

$$C_\tau \geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\log\det\left(I_M + \frac{\rho}{1+\frac{M}{T_\tau}}\frac{\bar{H}\bar{H}^*}{M}\right). \tag{42}$$

   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

$$\begin{aligned} C_\tau &\geqslant \mathrm{E}\,\frac{T-T_\tau}{T}\mathrm{tr}\,\log\left(I_M + \frac{\rho^2 T_\tau}{M}\frac{\bar{H}\bar{H}^*}{M}\right) \\ &\approx \frac{T-T_\tau}{T}\mathrm{E}\,\mathrm{tr}\,\frac{\rho^2 T_\tau \log e}{M}\cdot\frac{\bar{H}\bar{H}^*}{M} \\ &= \frac{NT_\tau(T-T_\tau)\log e}{MT}\rho^2. \end{aligned} \tag{43}$$

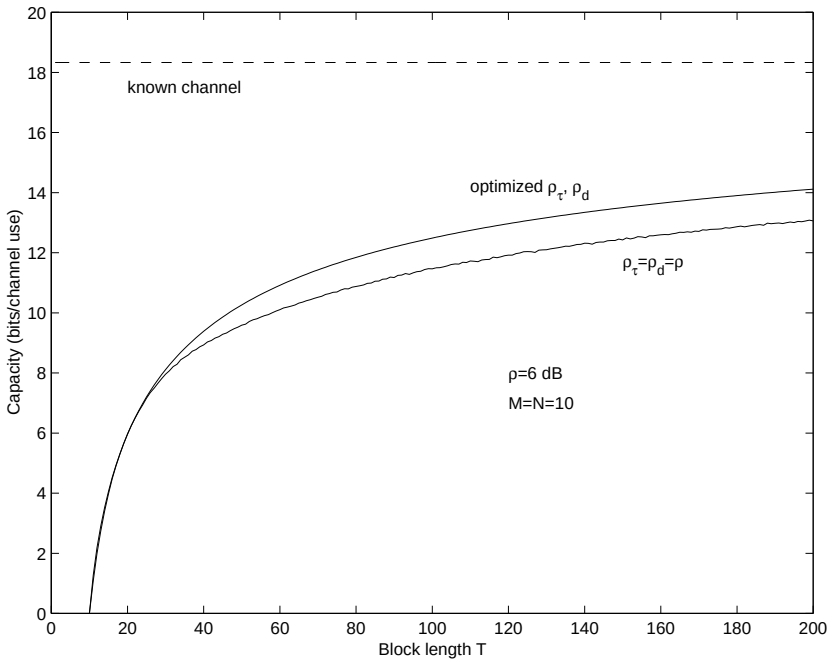

Figure 1: The training-based lower bound on capacity as a function of T when SNR $\rho = 6$ dB and $M = N = 10$, for optimized ρ_τ and ρ_d (upper solid curve, equation (33)) and for $\rho_\tau = \rho$ (lower solid curve, equation (41) optimized for T_τ). The dashed line is the capacity when the receiver knows the channel.

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

$$C_\tau \geqslant \frac{NT \log e}{4M} \rho^2. \tag{44}$$

This expression coincides with the expression obtained in Section 3.3.1. 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.

4 Plots of Training Intervals and Capacities

Figures 1 and 2 display the capacity obtained as a function of the blocklength 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 (33) in Theorem 3, and the curves for $\rho_\tau = \rho_d = \rho$ were obtained by maximizing (41) over T_τ.

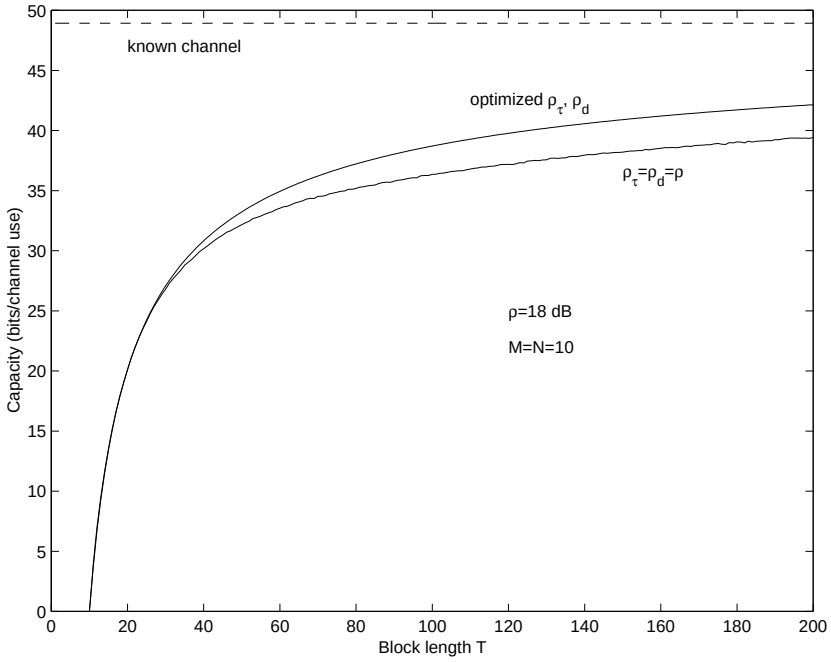

Figure 2: Same as Figure 1, except with $\rho = 18$ dB.

We know that if ρ_τ and ρ_d are optimized then the optimal training interval $T_\tau = M$, but when the constraint $\rho_\tau = \rho_d = \rho$ is imposed then $T_\tau \geqslant M$. Figure 3 displays the T_τ that maximizes (41) 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 3.4 that as $\rho \to 0$ the training increases until it reaches $T/2$.

Figure 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 optimal 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. Figures 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

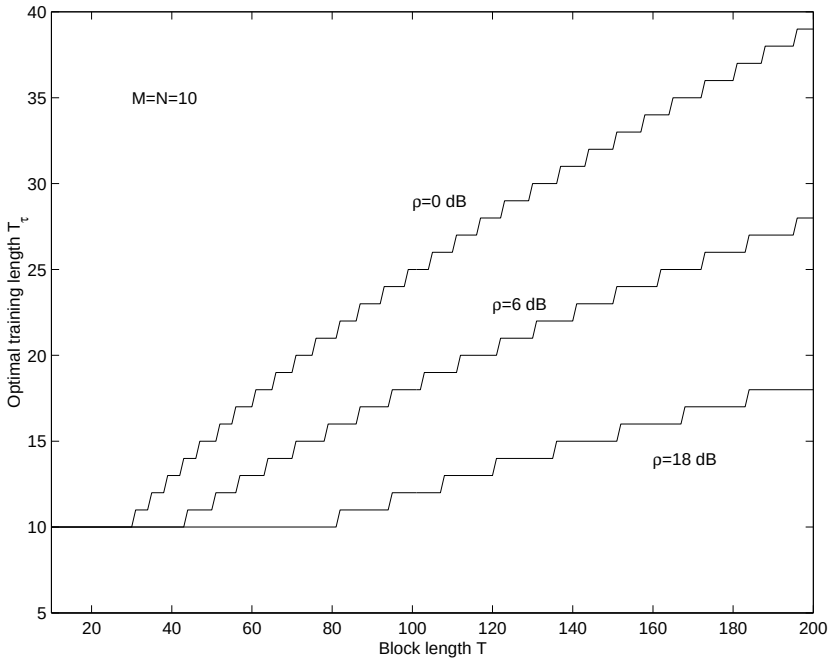

Figure 3: The optimal amount of training T_τ as a function of block length T for three different SNR's ρ, for $M = N = 10$ and constraining the training and data powers to be equal $\rho_\tau = \rho_d = \rho$. The curves were made by numerically finding the T_τ that maximized (41).

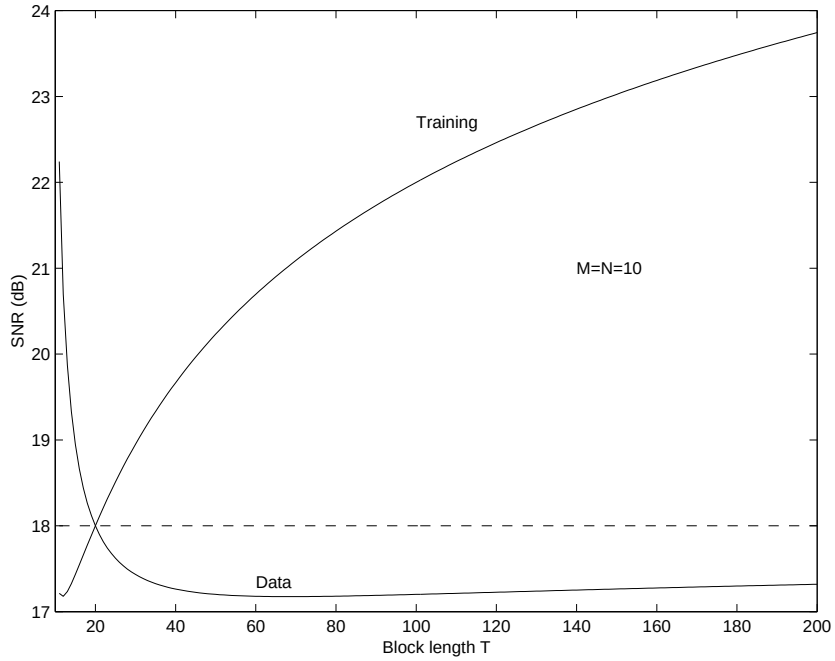

Figure 4: The optimal power allocation ρ_τ (training) and ρ_d (data transmission) as a function of block length T for $\rho = 18$ dB (shown in the dashed line) with $M = N = 10$. These curves are drawn from Theorem 2 and equations (28) for $T_\tau = M$.

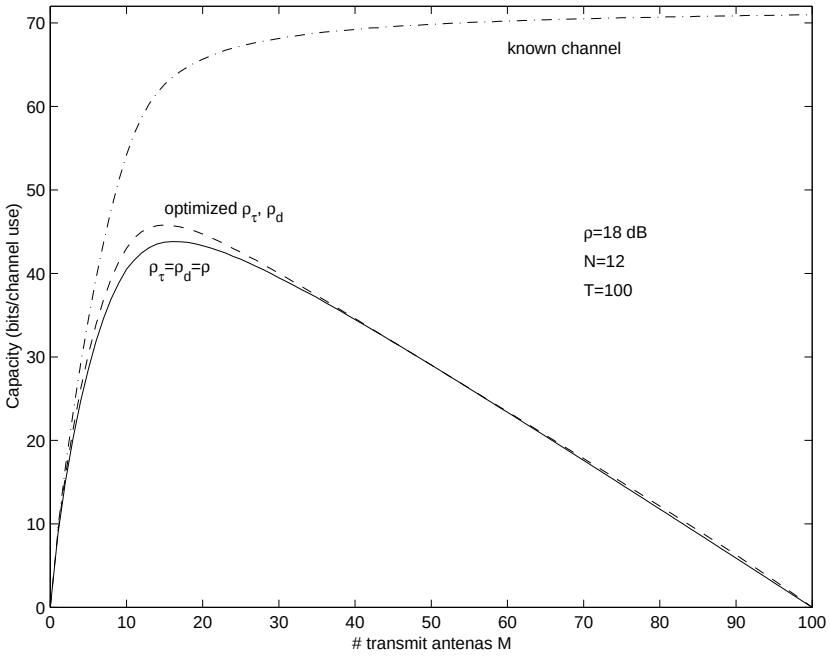

Figure 5: Capacity as a function of number of transmit antennas M with $\rho = 18$ dB and $N = 12$ receive antennas. The solid line is optimized over T_τ for $\rho_\tau = \rho_d = \rho$ (equation (41)), and the dashed line is optimized over the power allocation with $T_\tau = M$ (Theorem 3). The dash-dotted line is the capacity when the receiver knows the channel perfectly. The maximum throughput is attained at $M \approx 15$.

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.

5 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

$$C(\rho, T, M, N) \geqslant \left(1 - \frac{M}{T}\right) \mathrm{E} \log\det \left(I_M + \frac{1}{(\sqrt{1 - \frac{M}{T}} + \sqrt{\frac{M}{T}})^2} \rho \frac{\bar{H}\bar{H}^*}{M} \right), \tag{45}$$

where $\bar{H}$ 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.

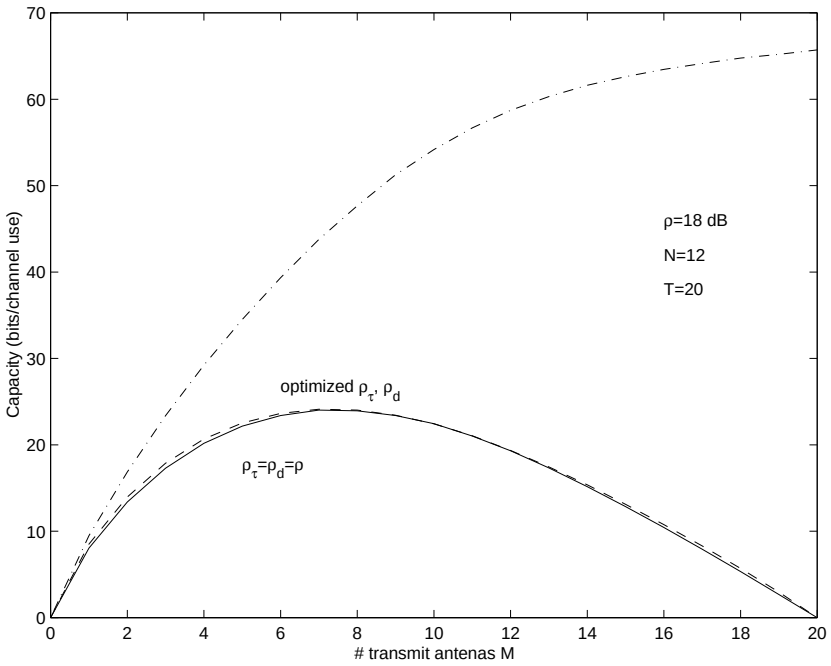

Figure 6: Same as Figure 5, except with $T = 20$. The maximum throughput is attained at $M \approx 7$. Observe that the difference between optimizing over ρ_τ and ρ_d versus setting $\rho_\tau = \rho_d = \rho$ is negligible.

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 are highly suboptimal when T is "close" to M. In fact, when $T = M$, the resulting capacity bound is zero since the training phase occupies the entire coherence interval. Figures 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 available but we elect to use only $M' \leqslant M$ of them in a training-based scheme. Equation (45) is then rewritten as

$$C(\rho, T, M, N) \geqslant \max_{M' \leqslant M} \left(1 - \frac{M'}{T}\right) \mathrm{E}\, \log\det \left(I'_M + \frac{1}{(\sqrt{1 - \frac{M'}{T}} + \sqrt{\frac{M'}{T}})^2} \rho \frac{\bar{H}\bar{H}^*}{M'} \right). \qquad (46)$$

Defining $Q = \min(M', N)$ and λ to be an arbitrary nonzero eigenvalue of $\frac{1}{(\sqrt{1-\frac{M'}{T}}+\sqrt{\frac{M'}{T}})^2} \frac{\bar{H}\bar{H}^*}{M'}$, 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

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

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,

$$C(\rho, T, M, N) \geqslant \left(1 - \frac{K}{T}\right) K \log \rho. \tag{47}$$

This argument implies that at high SNR the optimal number of transmit antennas to use in a training-based scheme is $K = \min(M, N, T/2)$. We argue in Section 3 that the whole process of training is highly suboptimal at low SNR. We now ask whether the same is true at high SNR, and whether our bounds are tight? The answer to this question can be found in the recent work [11] 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 are optimal, provided we use $K = \min(M, N, T/2)$ transmit antennas. (A similar conclusion is also drawn in [11]). We see indications of this result in Figure 5 where the maximum throughput is attained at $M \approx 15$ versus the predicted high SNR value of $K = 12$, and in Figure 6 at $M \approx 7$ versus the predicted $K = 10$.

We noted in the paragraph before Section 3.1 that our training-based capacity bounds are tight as $\rho \to 0$, since the additive noise term behaves as Gaussian noise at low SNR. The resulting training-based performance is extremely poor because the training-based capacity behaves like ρ^2, whereas the actual capacity decays as ρ. The exact transition between what should be considered "high" SNR where training yields 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 use training.

A Proof of Worst-Case Noise Theorem

Consider the matrix-valued additive noise known channel

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

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.,

$$\mathrm{E}\,\frac{1}{M} SS^* = 1 \quad \text{and} \quad \mathrm{E}\,\frac{1}{N} VV^* = 1. \tag{A.2}$$

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 (A.1) subject to the power constraints (A.2).

A.1 The additive Gaussian noise channel

We begin by computing the capacity of the channel (A.1) when V has a zero-mean complex Gaussian distribution with variance $R_V = \mathrm{E}\, V^* V$ (additive Gaussian noise channel). We generalize the arguments of [1, 2], which assume $R_V = I_N$, in a straightforward manner.

The capacity is the maximum, over all input distributions, of the mutual information between the received signal and known channel $\{X, H\}$ and the transmitted signal S. Thus,

$$\begin{aligned} I(X, H; S) &= I(X; S|H) + \underbrace{I(H; S)}_{=0} \\ &= h(X|H) - h(X|S, H), \end{aligned}$$

where $h(\cdot)$ is the entropy function. Now, $X|\{H, S\}$ is complex Gaussian with variance R_V, and $X|H$ has variance $R_V + \frac{\rho}{M} H^* R_S H$, where $R_S = \mathrm{E}\, S^* S$. Moreover, $h(X|H)$ is maximized when its distribution is Gaussian (which can always be achieved by making S Gaussian). Since $h(X|S, H)$ does not depend on the distribution of S, we conclude that choosing S Gaussian with an appropriate covariance achieves capacity,

$$C = \max_{p_S(\cdot), \mathrm{E}\, SS^* = M} I(X, H; S) = \max_{R_S, \mathrm{tr}\, R_S = M} \mathrm{E}\, \log\det \pi e \left(R_V + \frac{\rho}{M} H^* R_S H \right) - \log\det \pi e R_V.$$

Thus, the channel capacity is

$$C = \max_{R_S, \operatorname{tr} R_S = M} \mathrm{E} \log\det \left(I_N + \frac{\rho}{M} R_V^{-1} H^* R_S H \right). \tag{A.3}$$

A.2 Uncorrelated noise—proof of worst-case noise theorem

To obtain the worst-case noise distribution for V satisfying (A.2), we shall first solve a special case when the noise V and the signal S are uncorrelated:

$$\mathrm{E}\, S^* V = 0_{M \times N}. \tag{A.4}$$

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 yields

$$C_{\text{worst}} \leqslant \min_{R_V, \operatorname{tr} R_V = N} \max_{R_S, \operatorname{tr} R_S = M} \mathrm{E} \log\det \left(I_N + \frac{\rho}{M} R_V^{-1} H^* R_S H \right). \tag{A.5}$$

To obtain a lower bound on C_{worst}, we compute the mutual information for the channel (A.1) assuming that S is zero-mean complex Gaussian with covariance matrix R_S, but that the distribution on V is arbitrary. Thus,

$$I(X; S|H) = h(S|H) - h(S|X, H) = \log\det \pi e R_S - h(S|X, H).$$

Computing the conditional entropy $h(S|X, H)$ requires an explicit distribution on V. However, if the covariance matrix $\operatorname{cov}(S|X, H) = \mathrm{E}_{|X,H}(S - \mathrm{E}_{|X,H} S)^*(S - \mathrm{E}_{|X,H} S)$ of the random variable $S_{|X,H}$ is known, $h(S|X, H)$ has the upper bound

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

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 $\operatorname{cov}(S|X, H)$. Its proof can be found in, for example, [12].

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*

$$\mathit{cov}(S|X,H) = \mathrm{E}\,(S - \mathrm{E}_{|X,H}S)^*(S - \mathrm{E}_{|X,H}S) \leqslant \mathrm{E}\,(S-\hat{S})^*(S-\hat{S}). \tag{A.6}$$

Substituting the LMMSE (linear-minimum-mean-square-error) estimate $\hat{S} = XR_X^{-1}R_{XS}$ in this lemma yields

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

With the channel model (A.1)–(A.4), we see that

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

Thus,

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

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

$$I(X;S|H) \geqslant \mathrm{E}\,\log\det\left(I_N + \frac{\rho}{M}R_V^{-1}H^*R_SH\right)^{-1}. \tag{A.7}$$

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

$$C_{\mathrm{worst}} \geqslant \min_{R_V,\,\mathrm{tr}\,R_V=N}\ \max_{R_S,\,\mathrm{tr}\,R_S=M}\mathrm{E}\,\log\det\left(I_N + \frac{\rho}{M}R_V^{-1}H^*R_SH\right). \tag{A.8}$$

The combination of this inequality and (A.5) yields

$$C_{\mathrm{worst}} = \min_{R_V,\,\mathrm{tr}\,R_V=N}\ \max_{R_S,\,\mathrm{tr}\,R_S=M}\mathrm{E}\,\log\det\left(I_N + \frac{\rho}{M}R_V^{-1}H^*R_SH\right). \tag{A.9}$$

To prove the inequalities in (13), 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 (A.7), 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 $R_S = U\Lambda_s U^*$, 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 $R_S = \frac{1}{M!}\sum_{m=1}^{M!} P_m R_{S,\mathrm{opt}} P_m^* = I_M$, where the $P_1, \ldots, P_{M!}$ are all possible $M \times M$ permutation matrices. Since the "expected log-det" function in (A.9) is concave in R_S, 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 (A.9) is convex in R_V.

A.3 Correlated Noise

We can also find the worst case general additive noise, possibly correlated with the signal S. We do not use this result in the body of the paper because it is not always amenable to closed-form analysis. For simplicity, we assume a rotationally-invariant distribution for H.

Any arbitrary noise can be decomposed as

$$V = \underbrace{V - SR_S^{-1}R_{SV}}_{V'} + SR_S^{-1}R_{SV}, \tag{A.10}$$

where $V^{'}$ is uncorrelated with S. Thus, (A.1) can be written as

$$X = S\left(\sqrt{\frac{\rho}{M}}H + R_S^{-1}R_{SV}\right) + V^{'}.$$

Defining $A \triangleq \sqrt{M}R_S^{-1}R_{SV}$, we have

$$X = S\frac{\sqrt{\rho}H + A}{\sqrt{M}} + V^{'}, \tag{A.11}$$

where V' is uncorrelated with S and has the power constraint

$$\frac{1}{N}\mathrm{E}\, V^{'}V^{'*} = \frac{1}{N}\mathrm{E}\, VV^* - \frac{1}{MN}\mathrm{E}\, SAA^*S^* = 1 - \frac{1}{MN}\mathrm{tr}\, A^*R_S A = \sigma_{V'}^2.$$

The worst-case uncorrelated noise V' has therefore the distribution $\mathcal{CN}(0, \sigma_{V'}I_N)$, and the capacity for the

channel (A.11) becomes

$$\mathrm{E}\,\log\det\left(I_M+\frac{(\sqrt{\rho}H+A)(\sqrt{\rho}H+A)^*}{M\sigma_{V'}^2}\right).$$

Since the capacity-achieving distribution on S is $\mathcal{CN}(0,I_M)$, [1] we have $R_S = I_M$ and so $\sigma_{V'}^2 = 1 - \frac{1}{MN}\mathrm{tr}\,A^*A$, so that the capacity becomes

$$\mathrm{E}\,\log\det\left(I_M+\frac{(\sqrt{\rho}H+A)(\sqrt{\rho}H+A)^*}{M(1-\frac{1}{MN}\mathrm{tr}\,A^*A)}\right).$$

Clearly, the worst-case additive noise is found by minimizing the above expression over the matrix $A \in \mathcal{C}^{M\times N}$, subject to the constraint $\mathrm{tr}\,A^*A \leqslant MN$. Hence, we have shown the following result.

Theorem 4 (Worst-Case 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 with a rotationally-invariant distribution, 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

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

Then the worst-case noise is given by $V=\sqrt{\frac{1}{M}}SA+W$, where W is independent zero-mean Gaussian noise with variance $\sigma^2 = 1-\frac{1}{N}\mathrm{tr}\,AA^*$, i.e., $W \sim \mathcal{CN}(0,\sqrt{1-\frac{1}{N}\mathrm{tr}\,AA^*}I_N)$, and where $A \in \mathcal{C}^{M\times N}$ is the matrix solution to

$$C_{\mathrm{worst}} = \min_{\{A,\mathrm{tr}\,AA^*<MN\}} \mathrm{E}\,\log\det\left(I_M+\frac{(\sqrt{\rho}H+A)(\sqrt{\rho}H+A)^*}{M(1-\frac{1}{MN}\mathrm{tr}\,AA^*)}\right). \tag{A.12}$$

We also have the minimax property

$$I_{V\sim AS+\mathcal{CN}(0,\sigma^2 I_N),S}(X;S) \;\leqslant\; I_{V\sim AS+\mathcal{CN}(0,\sigma^2 I_N),S\sim\mathcal{CN}(0,I_M)}(X;S) = C_{\mathrm{worst}} \;\leqslant\; I_{V,S\sim\mathcal{CN}(0,I_M)}(X;S). \tag{A.13}$$

We do not know how to find an explicit solution to the optimization problem (A.12) in general. When the

[1] Recall that the transmitter has no knowledge of the channel H, and hence of the matrix A, so that it cannot minimize the noise power $\sigma_{V'}^2 = 1-\frac{1}{MN}\mathrm{tr}\,A^*R_SA$ by a clever choice of R_S—the best it can do is $R_S = I_M$.

channel is scalar, however, we can solve it easily.

Corollary 3 (Scalar Case). *Consider the scalar channel additive noise channel*

$$x = \sqrt{\rho}s + v,$$

where the signal s and the additive noise v satisfy the power constraints $\mathrm{E}\,|s|^2 = \mathrm{E}\,|v|^2 = 1$. Then the worst-case noise is given by $v = as + w$ where w is independent zero-mean Gaussian noise with variance $1 - |a|^2$ and where

$$a = \begin{cases} -\sqrt{\rho} & \text{if } \rho < 1 \\ -\sqrt{\frac{1}{\rho}} & \text{if } \rho \geqslant 1, \end{cases}$$

The resulting worst-case capacity is

$$C = \begin{cases} 0 & \text{if } \rho < 1 \\ \log \rho & \text{if } \rho \geqslant 1, \end{cases}$$

Note that, when $\rho < 1$, the noise has enough power to subtract out the effect of the signal so that the resulting capacity is zero. When $\rho > 1$, however, the noise only subtracts out a "portion" of the signal and reserves the remainder of its power for independent Gaussian noise. The resulting worst-case capacity is $\log \rho$, as compared with $\log(1 + \rho)$, the worst-case capacity with uncorrelated noise. Thus, at high SNR, correlated noise does not affect the capacity much more than uncorrelated noise.

References

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecom.*, 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 Annual Allerton Conference on Communications, Control, and Computing*, Sept. 22–24 1999.

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

[5] M. Medard, "The effect upon channel capacity in wireless communication of perfect and imperfect knowledge of the channel," *to appear in IEEE Trans. Info. Theory*.

[6] B. Hochwald and W. Sweldens, "Differential unitary space time modulation," tech. rep., Bell Laboratories, Lucent Technologies, Mar. 1999. To appear in *IEEE Trans. Comm.*. Download available at `http://mars.bell-labs.com`.

[7] B. Hughes, "Differential space-time modulation," *submitted to IEEE Trans. Info. Theory*, 1999.

[8] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity," *to appear in J. Sel. Area Comm.*, 2000.

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

[10] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time Rayleigh fading channels," in *IEEE Int. Symp. Info. Theory*, p. 473, June 1997. Also submitted to *IEEE Trans. Info. Theory*.

[11] L. Zheng and D. Tse, "Packing spheres in the Grassman manifold: a geometric approach to the noncoherent multi-antenna channel," *submitted to IEEE Trans. Info. Theory*, 2000.

[12] T. Söderström and P. Stoica, *System Identification*. London: Prentice Hall, 1989.