# On the Capacity of Frequency-Selective Channels in Training-Based Transmission Schemes

H. Vikalo[†], B. Hassibi[†], B. Hochwald[‡], and T. Kailath[§]

**Abstract**

Communication systems transmitting over frequency-selective channels generally employ an equalizer to recover the transmitted sequence corrupted by intersymbol interference. Most practical systems use a training sequence to learn the channel impulse response and thereby design the equalizer. An important issue is determining the optimal amount of training: too little training and the channel is not learned properly, too much training and there is not enough time available to transmit data before the channel changes and must be learned anew. We use an information-theoretic approach to find the optimal parameters in training-based transmission schemes for channels described by a block-fading model. The optimal length of the training interval is found by maximizing a lower bound on the training-based channel capacity. When the transmitter is capable of providing two distinct transmission power levels, one for training and one for data transmission, the optimal length of the training interval is shown to be equal to the length of the channel. Further, we show that, at high SNR, training-based schemes achieve the capacity of block-fading frequency selective channels, whereas at low SNR they are highly suboptimal.

[†]California Institute of Technology, Pasadena, CA 91125
[‡]Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974
[§]Information Systems Laboratory, Stanford University, Stanford, CA 94309

# 1 Introduction

Frequency selective fading multipath channels are often encountered in wireless communication systems (see the review of fading channels [1] and the references therein). To combat intersymbol interference (ISI) on such channels, receivers use various equalization techniques. Most practical communication systems learn the channel impulse response by means of training. They devote a portion of the transmission time to the training sequence, known to the receiver. This training sequence, and its transmitted version corrupted by ISI and additive noise which is received at the other end, are then used to estimate the channel. Alternatively, the channel may be identified blindly, using available information about the channel and input signals.

The most important parameter in any training-based scheme is the length of training interval or, equivalently, the amount of training. Clearly, too little training and the channel is not learned properly, too much training and there is not enough time to transmit information before the channel changes and must be relearned. One can conceive of different criteria to determine the optimal amount of training, but we believe that the most natural criterion is an information-theoretic one based on maximizing capacity. The reason being that in most communication systems one is concerned with reliably transmitting data at the highest possible rate, a concern that is well captured by capacity. Computing the capacity for frequency-selective fading channels in training-based schemes is a formidable task and so, as a compromise, we find a lower bound on this capacity. The channel is assumed to be discrete-time finite-impulse-response (FIR), subject to block-fading, i.e., the channel impulse response is constant over an interval (the so-called *coherence interval*) after which is changed to an independent value. The capacity lower bound is then maximized to find the optimal portion of the coherence interval that should be devoted to training. If the transmitter is capable of providing two distinct transmission power levels for training and for data transmission, then the optimal length of the training interval is shown to be equal to the length of the channel. If the transmitter cannot vary transmission power over the training and data transmission intervals, then the optimal length of the training interval can be found numerically, and may be longer than the length of the channel (up to one half of the coherence interval).

A natural question to ask is how good are training-based schemes? In other words, can one obtain significant improvement in performance by employing other methods, such as non-coherent detection or blind equalization? Our analysis allows us to give a qualitative answer to the above question. At high SNR, we show that our training-based capacity lower bounds coincide with the actual Shannon capacity of a block-fading FIR channel (this statement is true insofar as the leading order terms are concerned). Therefore, at high SNR, training-based schemes achieve the leading order term of the actual Shannon capacity and so we lose very little by considering a training-based scheme, as opposed to a more general scheme. On the other hand, at low SNR, we show that training-based schemes are highly suboptimal, in the sense that they can only deliver a diminishing factor of the actual Shannon capacity. Therefore, at low SNR, alternative methods to training must be sought.

Using a lower bound on the capacity of training-based schemes first developed in [2] and later extended to frequency-selective channels in [8], related work [9] finds power allocation and optimal placement for the training symbols. Similar techniques were also used for analysis of channel capacity in [3].

The remainder of the paper is organized as follows: In Section 2, we describe the block-fading model for FIR channels. The procedure for finding the optimal parameters of the transmission scheme is presented in Section 3. The optimization results are summarized in Theorem 1. In Section 4, we examine the performance of the optimal training-based transmission scheme at both high and low SNR by comparing the capacity bounds it achieves with the actual Shannon capacity of the block-fading frequency selective channel. This is illustrated by simulation results presented in Section 5.

## 2 Frequency-selective Channel Model

Frequency selective channels are characterized by a constant gain and linear phase response over a bandwidth which is smaller than the bandwidth of the signal to be transmitted. Equivalently, in the time domain, the length of the impulse response of the channel is equal to or longer than the width of the modulation signal. This is the case, for example, in high data rate wireless systems.

We shall assume a discrete-time block-fading frequency-selective FIR channel model $H(z) = h_1 + h_2 z^{-1} + \ldots + h_L z^{-L+1}$, where the channel coefficients $\{h_i\}_{i=1}^{L}$ are constant for some discrete interval of $T$ channel uses (referred to as the coherence interval), after which they change to independent values held constant for another coherence interval of length $T$, and so on. The block-fading model is appropriate for communication systems where, over a symbol block, the signaling rate is much faster than the pace at which the propagation environment changes. This is often the case in, e.g., TDMA or frequency-hopping-based systems.

We assume that the distribution of the coefficients making up the channel response is known at both the transmitter and receiver. During each coherence interval, to obtain the realization of the channel at the receiver, part of the interval is devoted to transmitting a known training sequence from which an estimate of the channel coefficients is obtained. Hence, in training-based transmission schemes, the coherence interval consists of the following two phases:

1. **Training Phase**

   During the training phase we transmit the $T_\tau$ training symbols $\theta_1, \ldots, \theta_{T_\tau}$. Since we are interested in estimating the $L$ channel coefficients $h_1, \ldots, h_L$, to obtain meaningful estimates we require $T_\tau \geq L$, which provides the receiver with at least as many equations as there are unknowns. To allow for the transmission of data, we clearly also require that $T_\tau < T$. If we collect the channel coefficients into the column vector $\mathbf{h} \in \mathcal{C}^{L\times 1}$, the received signals $y_1, \ldots, y_{T_\tau}$ into the column vector $\mathbf{y}_\tau \in \mathcal{C}^{T_\tau \times 1}$, and gather the training symbols into the $T_\tau \times L$ lower triangular Toeplitz matrix

$$\Theta_\tau = \begin{bmatrix} \theta_1 & 0 & 0 & \ldots & 0 \\ \theta_2 & \theta_1 & 0 & \ldots & 0 \\ \theta_3 & \theta_2 & \theta_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_L & \theta_{L-1} & \theta_{L-2} & \ldots & \theta_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{T_\tau} & \theta_{T_\tau - 1} & \theta_{T_\tau - 2} & \ldots & \theta_{T_\tau - L+1} \end{bmatrix}_{T_\tau \times L.}$$

   we may write

$$\mathbf{y}_\tau = \sigma_\tau \Theta_\tau \mathbf{h} + \mathbf{v}_\tau, \tag{1}$$

   where $\mathbf{v}_\tau \in \mathcal{C}^{T_\tau \times 1}$ is a vector of independent zero-mean unit variance additive complex Gaussian noise, and $\sigma_\tau^2$ is the expected transmit energy during the training phase. [In our scheme, the transmit powers during the training and data transmission phases may differ.] The training symbol vector $\theta_\tau = [\theta_1\ \theta_2\ \ldots\ \theta_{T_\tau}]^T$ satisfies the power constraint tr $\theta_\tau \theta_\tau^* = T_\tau$.

At the end of the training phase, using the observed signal $\mathbf{y}_\tau$ and the known training matrix $\Theta_\tau$, the receiver forms an estimate of the channel:

$$\hat{\mathbf{h}} = f(\mathbf{y}_\tau, \Theta_\tau). \tag{2}$$

Examples include the minimum-mean-square-error (MMSE) estimate

$$\hat{\mathbf{h}} = \sigma_\tau E\mathbf{h}\mathbf{h}^* \Theta_\tau^* \left(\sigma_\tau^2 \Theta_\tau E\mathbf{h}\mathbf{h}^* \Theta_\tau^* + I_{T_\tau}\right)^{-1} \mathbf{y}_\tau,$$

or the maximum-likelihood (ML)–equivalently, the zero-forcing (ZF)–estimate

$$\hat{\mathbf{h}} = \frac{1}{\sigma_\tau} E\mathbf{h}\mathbf{h}^* \Theta_\tau^* \left(\Theta_\tau E\mathbf{h}\mathbf{h}^* \Theta_\tau^*\right)^{-1} \mathbf{y}_\tau.$$

2. **Data Transmission Phase**

The date transmission phase consists of $T_d > 0$ channel uses. Collecting the transmitted symbols into the $T_d$-dimensional column vector $\mathbf{s}_d = [s_1\ s_2\ \ldots\ s_{T_d}]^T$, and the received signals into the $(T_d + L - 1)$-dimensional column vector $\mathbf{y}_d = [y_1\ y_2\ \ldots\ y_{T_d+L-1}]^T$, we may write

$$\mathbf{y}_d = \sigma_d H_d \mathbf{s}_d + \sigma_\tau H_\tau \theta_\tau + \mathbf{v}_d, \quad \mathrm{E}\mathbf{s}_d\mathbf{s}_d^* = R_\mathbf{s},\ \mathrm{tr} R_\mathbf{s} = T_d, \tag{3}$$

where $\mathbf{v}_d \in \mathcal{C}^{(T_\mathbf{d}+L-1)\times 1}$ is a vector of independent zero-mean unit variance additive complex Gaussian noise, $\sigma_d^2$ is expected transmit power during the data transmission phase, and where the Toeplitz matrices $H_d \in \mathcal{C}^{(T_d+L-1)\times T_d}$ and $H_\tau \in \mathcal{C}^{(T_d+L-1)\times T_\tau}$ are defined as

$$
H_d = \begin{bmatrix}
h_1 & & & & & & \\
h_2 & h_1 & & & & & \\
\vdots & \vdots & \ddots & & & & \\
h_L & h_{L-1} & \cdots & & h_1 & & \\
 & \ddots & & & & \ddots & \\
 & & & h_L & h_{L-1} & \cdots & h_1 \\
 & & & & \ddots & \vdots & \vdots \\
 & & & & & h_L & h_{L-1} \\
 & & & & & & h_L
\end{bmatrix}, \; H_\tau = \begin{bmatrix}
 & h_L & h_{L-1} & \cdots & h_2 \\
 & & h_L & \cdots & h_3 \\
 & & & \ddots & \vdots \\
 & & & & h_L \\
 & & & & \\
 & & & &
\end{bmatrix}.
$$

Note that, compared to the training phase, the roles of the transmitted symbols and channel coefficients are reversed: in the data transmission phase the former are the unknowns and the latter the knowns. Therefore, unlike (1), we have gathered the channel coefficients into a Toeplitz matrix and the transmitted symbols into a vector.

Two further points are worthy of mention. First, the term $H_\tau\theta_\tau$ in (3) represents the terms that "trickle down" from the training phase into the data transmission phase. Second, the receiver collects data even after the transmission ends and until the FIR channel is "emptied" (which is why the length of $\mathbf{y}_d$ is $T_d + L - 1$, rather than $T_d$).

Analysis of the transmission scheme requires knowledge of the covariance matrix of $\mathbf{h}$. We assume that the channel coefficients are independent complex Gaussian with covariance matrix

$$R_{\mathbf{h}} = E\mathbf{h}\mathbf{h}^* = \mathrm{diag}(\sigma^2_{h_1}, \sigma^2_{h_2}, \ldots, \sigma^2_{h_L}).$$

The estimate of the channel, $\hat{\mathbf{h}}$, is (indirectly, through $\hat{H}_d$ and $\hat{H}_\tau$) used to detect $\mathbf{s}_d$ from

$$\underbrace{\mathbf{y}_d - \sigma_\tau \hat{H}_\tau \theta_\tau}_{\mathbf{y}'_d} = \sigma_d \hat{H}_d \mathbf{s}_d + \underbrace{\sigma_d \tilde{H}_d \mathbf{s}_d + \sigma_\tau \tilde{H}_\tau \theta_\tau + \mathbf{v}_d}_{\mathbf{v}'_d}, \tag{4}$$

where $\mathbf{v}'_d$ is the effective noise comprised of the additive noise and residual channel estimation error, and $\mathbf{y}'_d$ denotes the combination of the measured and known signals during the data transmission phase.

We note that the following relations hold due to conservation of time and energy,

$$T = T_\tau + T_d, \quad \sigma^2 T = \sigma_\tau^2 T_\tau + \sigma_d^2 T_d,$$

where $\sigma^2 T$ is the total transmit energy, and $\sigma_\tau^2 T_\tau$ and $\sigma_d^2 T_d$ are the total transmit energies in the training and data phase, respectively.

Clearly, increasing $T_\tau$ improves the channel estimate, but that is achieved at the expense of the length of data transmission interval $T_d$. A similar trade-off holds between $\sigma_\tau^2$ and $\sigma_d^2$. We are interested in finding optimal values for the parameters $(T_\tau, T_d, \sigma_\tau^2, \sigma_d^2)$, along with the optimal training sequence $\theta_\tau$. Of course, we must specify what we mean by optimal. In the communications context we are studying, the most natural objective appears to be channel capacity, since it determines the maximum amount of information that can be reliably transmitted. Therefore in what follows we shall attempt to find the parameters $(T_\tau, T_d, \sigma_\tau^2, \sigma_d^2, \theta_\tau)$, that maximize the capacity of a training-based scheme.

In what follows we shall find it convenient to make use of the MMSE estimate of the channel $\mathbf{h}$. The MMSE estimate $\hat{\mathbf{h}}$ is the conditional mean of $\mathbf{h}$ given $\theta_\tau$ and $\mathbf{y}_\tau$, meaning that $\hat{\mathbf{h}}$ and $\tilde{\mathbf{h}}$, and so $\hat{H}$ and $\tilde{H}$, are uncorrelated. This implies that, when the channel estimate is MMSE, the effective noise $\mathbf{v}'_d$ in (4) is uncorrelated with the signal $\mathbf{s}_d$:

$$\begin{aligned} E\mathbf{s}_d {\mathbf{v}'_d}^* &= E\mathbf{s}_d \left(\sigma_d \tilde{H}_d \mathbf{s}_d + \sigma_\tau \tilde{H}_\tau \theta_\tau + \mathbf{v}_d\right)^* \\ &= \sigma_d E(\mathbf{s}_d \mathbf{s}_d^* \tilde{H}_d^*) + \sigma_\tau E(\mathbf{s}_d \theta_\tau^* \tilde{H}_\tau^*) + \underbrace{E(\mathbf{s}_d \mathbf{v}_d^*)}_{=0} \\ &= \sigma_d E(\mathbf{s}_d \mathbf{s}_d^*) E(\tilde{H}_d^*) + \sigma_\tau E(\mathbf{s}_d \theta_\tau^*) E(\tilde{H}_\tau^*) \\ &= 0 \end{aligned}$$

since $E(\mathbf{h} - \hat{\mathbf{h}}) = 0$ and, thus, $E(\tilde{H}_d) = E(H_d - \hat{H}_d) = 0$ and $E(\tilde{H}_\tau) = E(H_\tau - \hat{H}_\tau) = 0$. It is the property that $\mathbf{v}'_d$ and $\mathbf{s}_d$ are uncorrelated that makes the MMSE estimate useful in the sequel. We remark that no other estimate has this property.

# 3 A Lower Bound on the Training-Based Capacity

The capacity of any training-based scheme is given by the supremum of the mutual information between the transmitted signal $\mathbf{s}_d$ and the known and observed signals $\{\mathbf{y}_\tau, \theta_\tau, \mathbf{y}_d\}$. To be precise:

$$C_\tau = \sup_{p_{\mathbf{s}_d}, tr E\mathbf{s}_d\mathbf{s}_d^* \leq T_d, \theta_\tau} I(\mathbf{y}_\tau, \theta_\tau, \mathbf{y}_d; \mathbf{s}_d).$$

The above expression can be rewritten as

$$\begin{aligned} C_\tau &= \sup_{p_{\mathbf{s}_d}, tr E\mathbf{s}_d\mathbf{s}_d^* \leq T_d, \theta_\tau} \left[ I(\mathbf{y}_d; \mathbf{s}_d | \mathbf{y}_\tau, \theta_\tau) + \underbrace{I(\mathbf{y}_\tau, \theta_\tau; \mathbf{s}_d)}_{=0} \right] \\ &= \sup_{p_{\mathbf{s}_d}, tr E\mathbf{s}_d\mathbf{s}_d^* \leq T_d, \theta_\tau} I(\mathbf{y}_d; \mathbf{s}_d | \mathbf{y}_\tau, \theta_\tau) \end{aligned} \tag{5}$$

where we have used the fact that $\mathbf{s}_d$ and $\{\mathbf{y}_\tau, \theta_\tau\}$ are independent. This implies that the capacity in a training-based scheme is the supremum of the mutual information between the transmitted and received signals during the data transmission phase, given the transmitted and received signals during the training phase.

In general, finding the capacity in (5) is a hard problem. Instead, we find a lower bound on the capacity for a particular choice of the channel estimate. From (4),

$$\mathbf{y}'_d = \sigma_d \hat{H}_d \mathbf{s}_d + \mathbf{v}'_d. \tag{6}$$

We assume that $\hat{H}$ in (6) is the MMSE estimate of the channel $\mathbf{h}$. As mentioned earlier, this choice of the channel estimate guarantees that the signal $\mathbf{s}_d$ and the additive noise $\mathbf{v}'_d$ are uncorrelated. The reason explicit computation of the capacity in a training-based scheme is difficult is that the signals $\mathbf{s}_d$ and $\mathbf{v}'_d$, although uncorrelated, are dependent and that the distribution of $\mathbf{v}'_d$ is complicated to describe. Nonetheless, the covariance matrix of $\mathbf{v}'_d$ is easy to compute and has the following form

$$\begin{aligned} R_{\mathbf{v}'_d} &= \mathrm{E}\mathbf{v}'_d{\mathbf{v}'_d}^* = \mathrm{E}\left[\sigma_d^2 \tilde{H}_d \mathbf{s}_d \mathbf{s}_d^* \tilde{H}_d^* + \sigma_\tau^2 \tilde{H}_\tau \theta_\tau \theta_\tau^* \tilde{H}_\tau^* + \mathbf{v}_d \mathbf{v}_d^*\right] \\ &= \mathrm{E}\left[\sigma_d^2 \tilde{H}_d \mathbf{s}_d \mathbf{s}_d^* \tilde{H}_d^* + \sigma_\tau^2 \tilde{H}_\tau \theta_\tau \theta_\tau^* \tilde{H}_\tau^*\right] + I. \end{aligned} \tag{7}$$

In [2] it has been shown that among all additive noises that are uncorrelated with the signal $\mathbf{s}_d$ and have a given covariance matrix $R_{\mathbf{v}'_d}$, the worst-case noise (from a capacity point of view) is zero-mean independent Gaussian noise with the same covariance matrix. This allows us to lower bound the training-based capacity $C_\tau$, by the capacity of an additive Gaussian noise channel with the same noise covariance matrix. The capacity for such a channel, when the channel matrix is known at the receiver, has been computed in [4] (see also [5]) and so we may write

$$C_\tau \geq \max_{R_{\mathbf{s}_d}, tr R_{\mathbf{s}_d} = T_d, \theta_\tau} E \log\det\left(I + \sigma_d^2 R_{\mathbf{v}'_d}^{-1} \hat{H}_d R_{\mathbf{s}_d} \hat{H}_d^*\right),$$

where $R_{\mathbf{s}_d} = E\mathbf{s}_d\mathbf{s}_d^*$. Maximization of the above expression over $R_{\mathbf{s}_d}$ appears to be formidable. A choice that makes practical sense, since the transmitter does not know the channel, is $R_{\mathbf{s}_d} = I_{T_d}$. Since this choice will not necessarily maximize the right-hand-side of the above inequality, it yields the following lower bound

$$C_\tau \geq E \log\det\left(I + \sigma_d^2 R_{\mathbf{v}'_d}^{-1} \hat{H}_d \hat{H}_d^*\right).$$

Further, when $R_{\mathbf{s}} = I_{T_d}$ from (7) we obtain

$$R_{\mathbf{v}'_d} = I + \sigma_d^2 \underbrace{\mathrm{E}(\tilde{H}_d \tilde{H}_d^*)}_{R_{\mathbf{v}_1}} + \sigma_\tau^2 \underbrace{\mathrm{E}(\tilde{H}_\tau \theta_\tau \theta_\tau^* \tilde{H}_\tau^*)}_{R_{\mathbf{v}_2}} \tag{8}$$

It is also useful to define the normalized channel, $\bar{H}_d$, as

$$\bar{H}_d = \sqrt{\frac{LT_d}{\mathrm{trE}(\hat{H}_d^* \hat{H}_d)}} \hat{H}_d,$$

defined so that the nonzero entries of $\bar{H}_d$ have, on the average, unit variance. With this normalization we write capacity bound as

$$
\begin{aligned}
C_\tau &\geq E \log\det\left(I + \sigma_d^2 \frac{\mathrm{trE}(\hat{H}_d \hat{H}_d^*)}{LT_d} R_{\mathbf{v}'_d}^{-1} \bar{H}_d \bar{H}_d^*\right) \\
&= E \log\det\left(I + \sigma_d^2 \frac{T_d \sum_{i=1}^{L} \sigma_{\hat{h}_i}^2}{LT_d} R_{\mathbf{v}'_d}^{-1} \bar{H}_d \bar{H}_d^*\right),
\end{aligned} \tag{9}
$$

where $\sigma_{\hat{h}_i}^2 = E|\hat{h}_i|^2$. We are interested in finding the parameters of the transmission scheme such that the capacity lower bound in (9) is maximized. In particular, we will attempt to maximize this lower bound with respect to the training data sequence $\theta_\tau$, the training power $\sigma_d^2$, and the length of the training interval $T_\tau$.

# 4 The Optimal Training-Based Parameters

Maximization of the lower bound over the parameters $\{\theta_\tau, \sigma_d^2, T_\tau\}$ can be done in any order. We will find it most convenient to begin with determining the optimal training sequence $\theta_\tau$.

## 4.1 Maximizing over $\Theta_\tau$

From the orthogonality property of the MMSE estimate we may write

$$
\sigma_{\hat{h}_i}^2 = \sigma_{h_i}^2 - \sigma_{\tilde{h}_i}^2,
$$

where we have defined $\sigma_{\tilde{h}_i}^2 = E|\tilde{h}_i|^2 = E|h_i - \hat{h}_i|^2$. Thus (9) can be written as

$$
C_\tau \geq E \log\det\left(I + \sigma_d^2 \frac{\sum_{i=1}^{L} (\sigma_{h_i}^2 - \sigma_{\tilde{h}_i}^2)}{L} R_{\mathbf{v}'_d}^{-1} \bar{H}_d \bar{H}_d^*\right). \tag{10}
$$

Note that the RHS in (10) depends on $\theta_\tau$ through $\sigma_{\tilde{h}_i}^2$, $R_{\mathbf{v}'_d}$ and $\bar{H}_d$. However, we expect the dependence of $\bar{H}_d$ on $\theta_\tau$ to be the mildest, since the entries of $\bar{H}_d$ have been normalized. Exactly maximizing the RHS over $\theta_\tau$ appears to be analytically intractable, and so we propose to approximately maximize the RHS by considering only the dependence of $\sigma_{\tilde{h}_i}^2$ and $R_{\mathbf{v}'_d}$ of $\theta_\tau$, and ignoring the dependence on $\bar{H}_d$. In particular, we shall choose $\theta_\tau$ to maximize the effective power

$$\bar{\rho}_{\text{eff}} = \frac{\sigma_d^2}{L} \sum_{i=1}^{L} (\sigma_{h_i}^2 - \sigma_{\tilde{h}_i}^2). \tag{11}$$

From (11) this is equivalent to minimizing $\sum \sigma_{\tilde{h}_i}^2 = \text{tr} R_{\tilde{\mathbf{h}}}$. As we shall see below, this results in $R_{\mathbf{v}_2} = 0$ and also minimizes the trace of $R_{\mathbf{v}_1}$. Therefore it clearly minimizes the trace of $R_{\mathbf{v}'_d}$ and so the eigenvalues of $R_{\mathbf{v}'_d}^{-1}$, which is what appears in the capacity lower bound, will be large. Thus, we conclude that minimizing $\text{tr} R_{\tilde{\mathbf{h}}}$ makes very much sense.

To this end, consider the MMSE estimation error covariance matrix

$$\begin{aligned} R_{\tilde{\mathbf{h}}} &= R_{\mathbf{h}} - R_{\mathbf{hy}} R_{\mathbf{y}}^{-1} R_{\mathbf{yh}} \\ &= R_{\mathbf{h}} - R_{\mathbf{h}} (\sigma_\tau \Theta_\tau^*)(I + \sigma_\tau^2 \Theta_\tau R_{\mathbf{h}} \Theta_\tau^*)(\sigma_\tau \Theta_\tau) R_{\mathbf{h}} \\ &= (R_{\mathbf{h}}^{-1} + \sigma_\tau^2 \Theta_\tau^* \Theta_\tau)^{-1}. \end{aligned}$$

Hence, to minimize $\text{tr} R_{\tilde{\mathbf{h}}}$, we need to solve optimization problem

$$\min_{\theta_\tau, \text{tr}\theta_\tau \theta_\tau^* = T_\tau} \text{tr}(R_{\mathbf{h}}^{-1} + \sigma_\tau^2 \Theta_\tau^* \Theta_\tau)^{-1}. \tag{12}$$

The optimization problem (12) is convex and can be solved numerically as a semi-definite program. However, besides the optimal training sequence, in this paper we are interested in finding the optimal power allocation and the optimal length of the training interval. To carry out these optimization procedures, we need to obtain a closed-form expression of the training sequence. This can be facilitated if we assume that $\{h_1, \ldots, h_L\}$ are i.i.d. (that is, $R_{\mathbf{h}}$ is a multiple of the identity matrix).

Assume that $R_{\mathbf{h}} = I$. Then the optimization problem (12) can be expressed in terms of $\lambda_1, \lambda_2, \ldots, \lambda_L$, the eigenvalues of $\Theta_\tau^* \Theta_\tau$, as

$$\min_{\lambda_1, \ldots, \lambda_L} \sum_{i=1}^{L} \frac{1}{1 + \sigma_\tau^2 \lambda_i},$$

which is minimized when $\lambda_1 = \lambda_2 = \ldots = \lambda_L$ and, thus, when $\Theta_\tau^* \Theta_\tau$ is a multiple of identity. Since $\Theta$ has a Toeplitz structure, this can only be achieved by an impulse-like training sequence, where the impulse should be sent no later than $T_\tau - L + 1$ from the beginning of the training interval. To be more explicit, we have the following result.

**Optimal training sequence:** *In MMSE estimation of i.i.d. complex Gaussian channel coefficients, the training sequence that solves optimization problem (12) is given by*

$$\begin{aligned} \theta_{T_\tau - L + 2} &= \ldots = \theta_{T_\tau} = 0, \\ \theta_i &= \sqrt{T_\tau},\ \textit{for some } i,\ 1 \le i \le T_\tau - L + 1, \\ \theta_j &= 0\ \textit{for}\ 1 \le j \le T_\tau - L + 1, j \ne i. \end{aligned} \tag{13}$$

Note that the optimal training sequence is an impulse and that therefore the optimal $\Theta_\tau$ consists of a single diagonal. Since the duration of this diagonal is $L$, it may at first appear counterintuitive that such a diagonal matrix should be optimal for $T_\tau > L$, since in this case the matrix $\Theta_\tau$ will have $T_\tau - L$ zero rows. However, adding extra nonzero elements to $\Theta_\tau$ reduces the magnitude of these entries (since we have the constraint $\theta_\tau^* \theta_\tau = T_\tau$) and so the above calculations show that we do not gain by doing so.

As promised earlier, it follows that for the optimal choice of the training sequence $R_{\mathbf{v}_2} = 0$. This is a consequence of the readily verified fact that $\tilde{H}_\tau \theta_\tau = 0$. Moreover, since $\text{trace} R_{\mathbf{v}_1} = E\text{trace}\left(\tilde{H}_d \tilde{H}_d^*\right) = T_d \text{trace} R_{\tilde{\mathbf{h}}}$, we conclude that this choice of training sequence also minimizes the trace of $R_{\mathbf{v}'_d}$.

Inserting the optimal value of the training matrix $\Theta_\tau$ into (13) yields

$$R_{\tilde{\mathbf{h}}} = \frac{1}{1 + T_\tau \sigma_\tau^2} I, \tag{14}$$

or, in other words, $\sigma_{\tilde{h}_i}^2 = \sigma_{\tilde{h}}^2 = \frac{1}{1+T_\tau\sigma_\tau^2}$, for all $i$. Moreover, the matrix $R_{\mathbf{v}_1}$ in (8) is now diagonal,

$$R_{\mathbf{v}_1} = \text{diag}\left(\sigma_{\tilde{h}}^2, 2\sigma_{\tilde{h}}^2, \ldots, L\sigma_{\tilde{h}}^2, \ldots, L\sigma_{\tilde{h}}^2, \ldots, 2\sigma_{\tilde{h}}^2, \sigma_{\tilde{h}}^2\right).$$

It is therefore quite clear that

$$R_{\mathbf{v}_1} \le L\sigma_{\tilde{h}}^2 I_{T_d - L + 1},$$

so that $R_{\mathbf{v}'_d} \leq (1 + L\sigma_d^2\sigma_{\tilde{h}}^2)I$ and

$$R_{\mathbf{v}'_d}^{-1} \geq \frac{1}{1 + L\sigma_d^2\sigma_{\tilde{h}}^2} I_{T_d - L + 1}.$$

Now for positive definite matrices, $A \geq B > 0$, we have $\log\det A \geq \log\det B$.* Therefore we may write

$$E \log\det\left(I + \bar{\rho}_{\text{eff}} R_{\mathbf{v}'_d}^{-1} \bar{H}_d \bar{H}_d^*\right) \geq E \log\det\left(I + \frac{\bar{\rho}_{\text{eff}}}{1 + L\sigma_d^2\sigma_{\tilde{h}}^2} \bar{H}_d \bar{H}_d^*\right). \tag{15}$$

We thus obtain the following lower bound on the capacity of training-based schemes

$$C_\tau \geq E \log\det\left(I + \frac{\bar{\rho}_{\text{eff}}}{1 + L\sigma_d^2\sigma_{\tilde{h}}^2} \bar{H}_d \bar{H}_d^*\right).$$

Using $\sigma_{\tilde{h}}^2 = \frac{1}{1 + T_\tau \sigma_\tau^2}$ and $\bar{\rho}_{\text{eff}} = \sigma_d^2(1 - \sigma_{\tilde{h}}^2)$ allows us to write

$$C_\tau \geq E \log\det(I + \rho_{\text{eff}} \bar{H}_d \bar{H}_d^*), \tag{16}$$

where

$$\rho_{\text{eff}} = \frac{\bar{\rho}_{\text{eff}}}{1 + \sigma_d^2 \sum_{i=1}^{L} \sigma_{\tilde{h}_i}^2} = \frac{\sigma_d^2 \sigma_\tau^2 T_\tau}{1 + \sigma_\tau^2 T_\tau + \sigma_d^2 L}$$

can be treated as the effective SNR.

The expectation in (16) is taken over the random matrix $\bar{H}_d$, which is just a scaled version of $\tilde{H}_d = H_d - \hat{H}_d$. Note that the entries of $\tilde{H}_d$, the $\{\tilde{h}_i\}$, are iid $\mathcal{CN}(0, \sigma_{\tilde{h}}^2)$ random variables. Therefore the entries of $\bar{H}_d$, the $\{\bar{h}_i\}$ are iid $\mathcal{CN}(0, 1)$ random variables. This is a crucial result. The distribution of the elements of $\bar{H}_d$ does not depend on the training-based parameters $\{\sigma_\tau^2, \sigma_d^2, T_\tau, T_d\}$. Therefore one can ignore the distribution on $\bar{H}_d$ when maximizing the capacity lower bound over the remaining parameters.

The question that remains is how do the remaining parameters influence the capacity lower bound? The quantities $\sigma_\tau^2$ and $\sigma_d^2$ clearly only have an influence through $\rho_{\text{eff}}$. The parameters $T_\tau$ and $T_d$, on the other hand, influence both the effective SNR $\rho_{\text{eff}}$ *and* the dimension of the matrix $\bar{H}_d$, which is $(T_d + L - 1) \times T_d$.

Optimizing over the power allocations $\{\sigma_\tau^2, \sigma_d^2\}$ therefore appears to be easier and is what we do next.

* $A \geq B > 0$ implies that $B^{-*/2} A B^{1/2} \geq I$, which in turn implies that $\log\det B^{-*/2} A B^{1/2} \geq 0$ and which further implies that $\log\det A \geq \log\det B$.

## 4.2 Optimizing the Power Allocation

Clearly, maximizing the lower bound on the capacity over the power allocation is equivalent to maximizing $\rho_{\text{eff}}$ with respect to $\{\sigma_\tau^2, \sigma_d^2\}$. This is the following optimization problem

$$\max_{\sigma_d^2} \rho_{\text{eff}} \tag{17}$$

$$\text{s.t. } T_\tau + T_d = T, \quad \sigma_\tau^2 T_\tau + \sigma_d^2 T_d = \sigma^2 T.$$

Solving (17) is what we do in this section. To this end, let us denote the fraction of the total transmit energy used in the data transmission phase by $\alpha$, so that

$$\sigma_d^2 T_d = \alpha\sigma^2 T, \quad \sigma_\tau^2 T_\tau = (1-\alpha)\sigma^2 T, \quad 0 < \alpha < 1.$$

Then we can write

$$\begin{aligned}
\rho_{\text{eff}} &= \frac{\sigma_d^2 \sigma_\tau^2 T_\tau}{1 + \sigma_\tau^2 T_\tau + \sigma_d^2 L} \\
&= \frac{(\sigma^2 T)^2 \alpha(1-\alpha)}{T_d + (1-\alpha)\sigma^2 T T_d + \alpha\sigma^2 T L} \\
&= \frac{\sigma^2 T}{T_d - L} \cdot \frac{\alpha(1-\alpha)}{-\alpha + \gamma},
\end{aligned} \tag{18}$$

where

$$\gamma = \frac{T_d(1+\sigma^2 T)}{\sigma^2 T (T_d - L)}.$$

The maximum $\rho_{\text{eff}}$ in (18) depends on $\gamma$ and, consequently, the sign of $T_d - L$. We consider the three possible cases:

1. $T_d = L$:

   Here we can write

$$\rho_{\text{eff}} = \frac{(\sigma^2 T)^2}{L(1+\sigma^2 T)} \cdot \alpha(1-\alpha),$$

and, clearly,

$$\max_{\alpha} \rho_{\text{eff}} = \frac{(\sigma^2 T)^2}{4L(1+\sigma^2 T)}, \text{ achieved for } \alpha = \frac{1}{2}.$$

2. $T_d > L$:

   Differentiating $\rho_{\text{eff}}$ with respect to $\alpha$ yields

   $$\frac{d\rho_{\text{eff}}}{d\alpha} = \frac{\sigma^2 T}{T_d - L} \frac{(1-2\alpha)(-\alpha+\gamma)+\alpha(1-\alpha)}{(-\alpha+\gamma)^2}.$$

   Since $\gamma < 0$ here, the optimal $\alpha$ is given by

   $$\alpha = \gamma - \sqrt{\gamma(\gamma-1)},$$

   and the maximum effective power

   $$\max_{\alpha} \rho_{\text{eff}} = \frac{\sigma^2 T}{T_d - L}(\sqrt{\gamma} - \sqrt{\gamma - 1})^2.$$

3. $T_d < L$:

   In this case, $\gamma > 1$, and hence the optimal $\alpha$ is given by

   $$\alpha = \gamma + \sqrt{\gamma(\gamma-1)},$$

   achieving the maximum effective power

   $$\max_{\alpha} \rho_{\text{eff}} = \frac{\sigma^2 T}{L - T_d}(\sqrt{-\gamma} - \sqrt{-\gamma+1})^2.$$

We can summarize the previous results as follows:

**Optimal power distribution:** *Given the total transmit energy in a training-based transmission scheme, $\sigma^2 T$, the power allocation to the data transmission over an interval $T_d$ that maximizes the lower bound on* *the capacity of training-based schemes (16) is given by*

$$\sigma_d^2 = \begin{cases} (\gamma - \sqrt{\gamma(\gamma-1)})\sigma^2 \frac{T}{T_d} & \textit{for } T_d > L \\ \frac{1}{2}\sigma^2 \frac{T}{T_d} & \textit{for } T_d = L \\ (\gamma + \sqrt{\gamma(\gamma-1)})\sigma^2 \frac{T}{T_d} & \textit{for } T_d < L \end{cases}$$

*where* $\gamma = \frac{T_d(1+\sigma^2 T)}{\sigma^2 T(T_d - L)}$*. The corresponding lower bound on capacity is*

$$C_\tau \geq E \log\det(I + \rho_{\mathit{eff}} \bar{H}_d \bar{H}_d^*),$$

*where*

$$\rho_{\mathit{eff}} = \begin{cases} \frac{\sigma^2 T}{T_d - L}(\sqrt{\gamma} - \sqrt{\gamma - 1})^2 & \textit{for } T_d > L \\ \frac{(\sigma^2 T)^2}{4L(1+\sigma^2 T)} & \textit{for } T_d = L \\ \frac{\sigma^2 T}{L - T_d}(\sqrt{-\gamma} - \sqrt{-\gamma + 1})^2 & \textit{for } T_d < L \end{cases} \tag{19}$$

We can further simplify these expressions for the special cases of low and high SNR. In particular, at high SNR we have

$$\lim_{\sigma^2 \to \infty} \gamma = \frac{T_d}{T_d - L},$$

and at low SNR

$$\gamma = \frac{T_d}{T(T_d - L)} O(\frac{1}{\sigma^2}), \text{ as } \sigma^2 \to 0.$$

1. *High SNR:*

   At high SNR, the lower bound on capacity is given by

$$C_\tau \geq E \log\det(I + \frac{\sigma^2 T}{(\sqrt{T_d} + \sqrt{L})^2} \bar{H}_d \bar{H}_d^*),$$

   while the optimal power allocation is

$$\sigma_d^2 = \frac{\sqrt{T_d}}{\sqrt{T_d} + \sqrt{L}} \cdot \sigma^2 \frac{T}{T_d}.$$

2. *Low SNR:*

At low SNR, the optimal power allocation is

$$\sigma_d^2 = \frac{1}{2} \cdot \sigma^2 \frac{T}{T_d},$$

while the lower bound on the capacity is given by

$$\begin{aligned} C_\tau &\geq E \quad \log\det\left(I + \frac{(\sigma^2 T)^2}{4T_d}\bar{H}_d\bar{H}_d^*\right) \\ &\approx \quad \frac{(\sigma^2 T)^2}{4T_d} E\mathrm{trace}\left(\bar{H}_d\bar{H}_d^*\right) = \frac{(\sigma^2 T)^2 L}{4}, \end{aligned}$$

where to find the low SNR approximations, we used $\log\det(\cdot) = \mathrm{trace}\log(\cdot)$ and the Taylor series expansion of $\log(\cdot)$. Note that the low SNR capacity is quadratic in $\sigma^2$ and independent of $T_d$.

### 4.3 Optimizing over $T_\tau$

The final step in the optimization procedure is to maximize the capacity bound in (16) over the length of the training interval $T_\tau$ (equivalently, $T_d = T - T_\tau$). To this end, let us write the training-based capacity lower bound as

$$C_B(\rho_{\text{eff}}(T_d), T_d) = E\log\det\left(I + \rho_{\text{eff}}(T_d)\bar{H}_d^*\bar{H}_d\right) = E\log\det R(\rho_{\text{eff}}(T_d), T_d),$$

where, from the Toeplitz structure of $\bar{H}_d$, $R(\rho_{\text{eff}}(T_d), T_d)$ is the $T_d \times T_d$ Toeplitz matrix given by

$$R(\rho_{\text{eff}}(T_d), T_d)_{ij} = \begin{cases} 1 + \rho_{\text{eff}}(T_d)\sum_k |h_k|^2 & i = j \\ \rho_{\text{eff}}(T_d)\sum_k h_{k+i-j}^* h_k & i \neq j \end{cases}$$

Note that $T_d$ affects the capacity lower bound $C_B$ in two ways: one through the effective SNR, $\rho_{\text{eff}}$, and the other through the dimension of the Toeplitz matrix $R$. This is why we have explicitly written $C_B = C_B(\rho_{\text{eff}}(T_d), T_d)$.

For brevity, we shall here treat only the case when $T_d > L$. (Other cases are treated similarly.) Using some cumbersome algebra, it can be shown that the effective SNR, as given by (19) satisfies

$$\rho_{\text{eff}}(T_d) \geq \left(1 - \frac{1}{T_d}\right) \rho_{\text{eff}}(T_d - 1). \tag{20}$$

Thus, the rate of decrease of the SNR as we increase the training interval is no greater than $1/T_d$. Now note that

$$\begin{aligned}
C_B(\rho_{\text{eff}}(T_d), T_d) &= E \log\det\left(I + \rho_{\text{eff}}(T_d)\bar{H}_d^* \bar{H}_d\right) \\
&\geq E \log\det\left(I + (1 - \frac{1}{T_d})\rho_{\text{eff}}(T_d - 1)\bar{H}_d^* \bar{H}_d\right) \\
&= C_B(\rho_{\text{eff}}(T_d - 1), T_d) + E \log\det\left(I - \frac{\rho_{\text{eff}}(T_d - 1)}{T_d}\bar{H}_d^* \bar{H}_d R^{-1}(\rho_{\text{eff}}(T_d - 1), T_d)\right) \\
&= C_B(\rho_{\text{eff}}(T_d - 1), T_d) + E \log\det\left[I - \frac{1}{T_d}\left(I + \frac{1}{\rho_{\text{eff}}(T_d - 1)}(\bar{H}_d^* \bar{H}_d)^{-1}\right)^{-1}\right] \\
&= C_B(\rho_{\text{eff}}(T_d - 1), T_d) + E \sum_{i=1}^{T_d} \log\left(1 - \frac{1/T_d}{1 + \frac{1}{\lambda_i(\bar{H}_d^* \bar{H}_d)\rho_{\text{eff}}(T_d - 1)}}\right) \\
&= C_B(\rho_{\text{eff}}(T_d - 1), T_d) + E \sum_{i=1}^{T_d} \log \frac{1 + (1 - 1/T_d)\lambda_i(\bar{H}_d^* \bar{H}_d)\rho_{\text{eff}}(T_d - 1)}{1 + \lambda_i(\bar{H}_d^* \bar{H}_d)\rho_{\text{eff}}(T_d - 1)},
\end{aligned}$$

where $\lambda_i(\bar{H}_d^* \bar{H}_d)$ denote eigenvalues of $\bar{H}_d^* \bar{H}_d$.

Let us now further identify the first term on the RHS of the above inequality. Note that by the Toeplitz property of $R(\rho_{\text{eff}}(T_d - 1), T_d)$ we may write

$$R(\rho_{\text{eff}}(T_d - 1), T_d) = \begin{bmatrix} R(\rho_{\text{eff}}(T_d - 1), T_d - 1) & \times \\ \times & \times \end{bmatrix},$$

and that

$$\det R(\rho_{\text{eff}}(T_d - 1), T_d) = \frac{\det R(\rho_{\text{eff}}(T_d - 1), T_d - 1)}{[R^{-1}(\rho_{\text{eff}}(T_d - 1), T_d)]_{T_d, T_d}},$$

since $1/\left[R^{-1}(\rho_{\text{eff}}(T_d - 1), T_d)\right]_{T_d, T_d}$ is the corresponding Schur complement. Inserting this last expression into the bound for $C_B(\rho_{\text{eff}}(T_d), T_d)$ yields

$$C_B(\rho_{\text{eff}}(T_d), T_d) \geq C_B(\rho_{\text{eff}}(T_d-1), T_d-1) + \delta, \tag{21}$$

where

$$\delta = -E\log\left[R^{-1}(\rho_{\text{eff}}(T_d-1), T_d)\right]_{T_d,T_d} + E\sum_{i=1}^{T_d}\log\frac{1+(1-1/T_d)\lambda_i(\bar{H}_d^*\bar{H}_d)\rho_{\text{eff}}(T_d-1)}{1+\lambda_i(\bar{H}_d^*\bar{H}_d)\rho_{\text{eff}}(T_d-1)} \tag{22}$$

The quantity $C_B(\rho_{\text{eff}}(T_d), T_d)$ is the training-based capacity lower bound for a data interval of length $T_d$ and, likewise, $C_B(\rho_{\text{eff}}(T_d-1), T_d-1)$ is the training-based capacity lower bound for a data interval of length $T_d - 1$. Clearly, if $\delta \geq 0$ for all $T_d$, then the capacity lower bound is an increasing function of $T_d$ and so the optimal choice of the data interval is its largest possible value, $T_d = T - L$.

This is certainly true at high SNR. To see that, note that the second term in (22) is bounded below by

$$E\sum_{i=1}^{T_d}\log\frac{1+(1-1/T_d)\lambda_i(\bar{H}_d^*\bar{H}_d)\rho_{\text{eff}}(T_d-1)}{1+\lambda_i(\bar{H}_d^*\bar{H}_d)\rho_{\text{eff}}(T_d-1)} \quad \geq \quad \sum_{i=1}^{T_d}\log(1-\frac{1}{T_d}) = T_d\log(1-\frac{1}{T_d}),$$

which is increasing in $T_d$ and thus, since $T_d > L \geq 1$, is bounded below by $-2\log 2$. On the other hand, the first term in (22) behaves as $\log\sigma^2$ at high SNRs. We therefore have the following result.

**Optimal training interval at high SNR:** *For any $T$ and sufficiently high transmit power $\sigma^2$, the optimal length of the training interval is equal to the length of the channel, $T_\tau = L$.*

Showing that the above result holds for all SNR, requires one to verify that $\delta \geq 0$ at all SNR. Based on extensive simulations, we conjecture that the expected value of the first diagonal entry of $R^{-1}(\rho_{\text{eff}}(T_d - 1), T_d)$ is bounded by

$$E\log\left[R^{-1}(\rho_{\text{eff}}(T_d-1), T_d)\right]_{T_d,T_d} \leq E\sum_{i=1}^{T_d}\log\frac{1+(1-1/T_d)\lambda_i(\bar{H}_d^*\bar{H}_d)\rho_{\text{eff}}(T_d-1)}{1+\lambda_i(\bar{H}_d^*\bar{H}_d)\rho_{\text{eff}}(T_d-1)}.$$

We therefore propose the following.

**Conjecture:** *For any $T$ and any transmit power $\sigma^2$, the optimal length of the training interval is equal to the length of the channel, $T_\tau = L$.*

Although we have not yet been able to prove this conjecture, based on extensive study and simulation, we believe it to be true. We therefore formalize the result in the following theorem.

**Theorem 1 (Optimal Parameters of Training-Based Scheme).** *The optimal length of the training interval for a training-based transmission scheme over a block-fading frequency-selective FIR channel of length $L$ is given by $T_\tau = L$, and the lower bound on the capacity is given by*

$$C_\tau \geq E \log\det(I + \rho_{\mathit{eff}} \bar{H}^* \bar{H}),$$

*where*

$$\rho_{\mathit{eff}} = \begin{cases} \frac{\sigma^2 T}{T-2L}(\sqrt{\gamma} - \sqrt{\gamma - 1})^2 & \textit{for } T > 2L \\ \frac{(\sigma^2 T)^2}{4L(1+\sigma^2 T)} & \textit{for } T = 2L \\ \frac{\sigma^2 T}{2L-T}(\sqrt{-\gamma} - \sqrt{-\gamma+1})^2 & \textit{for } T < 2L \end{cases}$$

*and* $\gamma = \frac{(T-L)(1+\sigma^2 T)}{\sigma^2 T(T-2L)}$.

*The optimal power allocation is given by*

$$\sigma_d^2 = \begin{cases} (\gamma - \sqrt{\gamma(\gamma-1)})\sigma^2 \frac{T}{T-L} & \textit{for } T > 2L \\ \frac{1}{2}\sigma^2 \frac{T}{T-L} & \textit{for } T = 2L \\ (\gamma + \sqrt{\gamma(\gamma-1)})\sigma^2 \frac{T}{T-L} & \textit{for } T < 2L \end{cases},$$

$$\sigma_\tau^2 = \sigma_d^2 + (\sigma^2 - \sigma_d^2)\frac{T}{L}.$$

*At high SNR, the optimal power distribution can be approximated by*

$$\sigma_d^2 = \frac{T}{T - L + \sqrt{L(T-L)}}\sigma^2, \quad \sigma_\tau^2 = \frac{T}{L + \sqrt{L(T-L)}}\sigma^2,$$

*and the corresponding effective SNR becomes*

$$\rho_{\mathit{eff}} = \frac{\sigma^2 T}{(\sqrt{T-L} + \sqrt{L})^2}.$$

Some comments regarding Theorem 1 are in place. Intuitively, the longer the training interval, the better estimate of the channel, and thus the larger the effective SNR. However, a longer training interval means less time for the data transmission. Theorem 1 implies that the significance of the data transmission interval outweighs that of the training interval, as the optimal training interval is set to its minimum meaningful length–that is, the length of the channel. Although it would appear that this stems from the fact that increasing the length of the training interval is associated with the tendencies of logarithmic decrease and linear increase of the capacity lower bound, we should re-emphasize that this result is valid only when the optimal power allocation has been performed and not, for example, when $\sigma_\tau^2 = \sigma_d^2 = \sigma^2$.

From Theorem 1, we can draw some further conclusions about the general behavior of the power allocation. In particular, one can show that the following inequalities hold for all SNR $\sigma^2$

$$\begin{aligned}
\sigma_d^2 < \sigma^2 < \sigma_\tau^2, &\quad \text{for } T > 2L, \\
\sigma_\tau^2 < \sigma^2 < \sigma_d^2, &\quad \text{for } T < 2L, \\
\sigma_d^2 = \sigma^2 = \sigma_\tau^2, &\quad \text{for } T = 2L.
\end{aligned}$$

Hence, we need to spend more power for training than for transmission when $T > 2L$, more power for transmission than for training when $T < 2L$, and the same power for both when $T = 2L$.

### 4.4 Equal powers

The assumption made throughout the paper is that the communication system can provide two different transmission power levels, which are then allocated to the training and data transmission phases. However, this may not always be the case. If the practical constraints impose equal power in the training and transmission phases, i.e., if $\sigma_\tau^2 = \sigma_d^2 = \sigma^2$, then the capacity lower bound of (16) can be expressed as

$$C_\tau \geq E \log\det\left(I + \frac{\sigma^4 T_\tau}{1 + \sigma^2(T_\tau + L)} \bar{H}_d^* \bar{H}_d\right).$$

The trade-off between the training interval $T_\tau$ and the data transmission interval $T_d$ here is obvious. The capacity lower bound increases logarithmically with $T_\tau$ through $\log\det(\cdot)$ but decreases linearly through the dimension of $\bar{H}_d^* \bar{H}_d$.

Further simplifications of the capacity lower bound expressions are possible for the special cases of high and low SNR.

1. *High SNR:*

   At high SNR, we can write the capacity lower bound as

$$C_\tau \geq E \log\det\left(I + \frac{\sigma^2 T_\tau}{T_\tau + L} \bar{H}_d^* \bar{H}_d\right). \tag{23}$$

   The optimum length of the training interval can be obtained by evaluating (23) for various $T_\tau$, $L \leq T_\tau < T$.

2. *Low SNR:*

   At low SNR, using $\log(I + A) = A - A^2/2 + A^3/3 \ldots$, we obtain following expression for the capacity lower bound

$$\begin{aligned} C_\tau &\geq E \log\det\left(I + \sigma^4 T_\tau \bar{H}_d^* \bar{H}_d\right) \\ &\approx \mathrm{Etr}\sigma^4 T_\tau \bar{H}_d^* \bar{H}_d \\ &= \sigma^4 L T_\tau (T - T_\tau). \end{aligned}$$

   Upon taking the derivative with respect to $T_\tau$, one can notice that the capacity bound is maximized for $T_\tau = T/2$, i.e., that half of the coherence interval must be devoted to training. This matches the multi-antenna results of [2].

# 5 Comparing Capacity Bounds with Actual Capacity at High and Low SNR

Having determined the optimal amount of training for any training-based communication system, the question that remains is *how good are training-based schemes*? To answer this, one would need to compute the actual capacity of a block-fading frequency-selective channel and to compare it with the training-based capacity lower bounds that we have obtained. Unfortunately, computing this capacity, in the general case, is an open problem. However, we can obtain the expression for the capacity at high and low SNR ($\sigma \to \infty$ and $\sigma \to 0$).

We begin with the high SNR result, for which the following lemma is useful.

**Lemma 1.** *The* $(T+L-1) \times L$ *matrix*

$$S = \begin{bmatrix} s_1 & 0 & 0 & \dots & 0 \\ s_2 & s_1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_L & s_{L-1} & s_{L-2} & \cdots & s_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_T & s_{T-1} & s_{T-2} & \cdots & s_{T-L+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & s_T \end{bmatrix},$$

*is rank-deficient if, and only if,* $s_1 = s_2 = \ldots = s_T = 0$.

**Proof:** For any $L$ we prove the theorem by induction on $T$. Clearly, for $T = 1$ we have $S = s_1 I_L$, so that $S$ drops rank if, and only if, $s_1 = 0$. Suppose now that the result is true for $T-1$ and consider the case for $T$. Clearly, if $s_1 = s_2 = \ldots = s_T = 0$, we have $S = 0$ and so $S$ does not have full rank. Suppose now that $S$ is rank-deficient, which means that there exists an $L$-dimensional column vector $x$ such that $Sx = 0$. Focusing on the first $L$ equations this implies that

$$\begin{bmatrix} s_1 & 0 & 0 & \dots & 0 \\ s_2 & s_1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_L & s_{L-1} & s_{L-2} & \dots & s_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix} = 0,$$

which implies that we must have $s_1 = 0$. But this now implies that our problem is reduced to the rank-deficiency of the $(T+L-2) \times L$ lower triangular Toeplitz matrix with entries $s_2, \ldots, s_T$, for which we know that rank-deficiency occurs iff $s_2 = \ldots = s_T = 0$.

$\square$

We can now state the high SNR result.

**Theorem 2 (Capacity of an FIR Channel at High SNR).** *Consider the block-fading FIR channel model*

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{T+L-1} \end{bmatrix}}_{\mathbf{y}} = \sigma \underbrace{\begin{bmatrix} h_1 & & & & & & \\ h_2 & h_1 & & & & & \\ \vdots & \vdots & \ddots & & & & \\ h_L & h_{L-1} & \ldots & & h_1 & & \\ & \ddots & & & & \ddots & \\ & & & h_L & h_{L-1} & \ldots & h_1 \\ & & & & \ddots & \vdots & \vdots \\ & & & & & h_L & h_{L-1} \\ & & & & & & h_L \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_T \end{bmatrix}}_{\mathbf{s}} + \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{T+L-1} \end{bmatrix}}_{\mathbf{v}}, \tag{24}$$

*where the channel coefficients are constant for a coherence interval of $T$ channel uses, after which they change to independent values. The channel coefficients and the components of the noise vector $v$ are unknown to the transmitter and receiver and are iid $\mathcal{CN}(0,1)$ random variables. Then, when $T > L$, the high SNR ($\sigma \to \infty$) capacity of this channel, over $T$ channel uses, is given by*

$$C = (T - L)\log \sigma^2 + O(1). \tag{25}$$

**Proof:** In addition to writing the channel equation as $\mathbf{y} = \sigma H \mathbf{s} + \mathbf{v}$, we will also find it useful to rewrite it as

$$\mathbf{y} = \sigma S \mathbf{h} + \mathbf{v},$$

where $S$ is as in Lemma 1 and $\mathbf{h}$ is the $L$-dimensional column vector formed from the channel coefficients.

For reasons to be made clear shortly, we will need to distinguish between two cases: the case when $S$ is nonsingular (and hence by Lemma 1, $S \neq 0$), and the case when $S$ is singular (and hence by Lemma 1, $S = 0$). Thus, let $t$ denote a random variable that is one when $S \neq 0$ and is zero when $S = 0$. Moreover, assume that

$$p(t=0) = q \quad \text{and} \quad p(t=1) = 1-q.$$

We start by looking at the mutual information

$$I(\mathbf{y};\mathbf{s}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{s}). \tag{26}$$

Let us first focus on $h(\mathbf{y})$. Note that

$$\begin{aligned} h(\mathbf{y}) \leq h(\mathbf{y},t) &= h(\mathbf{y}|t) + h(t) \\ &= qh(\mathbf{y}|S=0) + (1-q)h(\mathbf{y}|S\neq 0) - q\log q - (1-q)\log(1-q) \\ &= qh(\mathbf{y}|S=0) + (1-q)h(\mathbf{y}|S\neq 0) + O(1). \end{aligned} \tag{27}$$

Now $\mathbf{y}|_{S=0}$ is just the random vector $\mathbf{v}$, which does not depend on $\sigma$, so that $h(\mathbf{y}|S=0) = O(1)$. Therefore we have

$$h(\mathbf{y}) \leq (1-q)h(\mathbf{y}|S\neq 0) + O(1),$$

and so we need to compute $h(\mathbf{y}|S \neq 0)$. To this end, let us partition (24) as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \sigma \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \mathbf{s} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}, \tag{28}$$

where $\mathbf{y}_1, \mathbf{s}, \mathbf{v}_1 \in \mathcal{C}^{T\times 1}$, $\mathbf{y}_2, \mathbf{v}_2 \in \mathcal{C}^{(L-1)\times 1}$, $H_1 \in \mathcal{C}^{T\times T}$, and $H_2 \in \mathcal{C}^{(L-1)\times T}$. Due to our statistical assumption on the channel coefficients, the matrix $H_1$ is generically invertible and so we may write

$$\mathbf{y}_2 = H_2 H_1^{-1}\mathbf{y}_1 + \mathbf{v}_2 - H_2 H_1^{-1}\mathbf{v}_1.$$

Now $h(\mathbf{y}|S \neq 0) = h(\mathbf{y}_1, \mathbf{y}_2|S \neq 0) = h(\mathbf{y}_1|S \neq 0) + h(\mathbf{y}_2|\mathbf{y}_1, S \neq 0)$, and the above relation clearly shows that $h(\mathbf{y}_2|\mathbf{y}_1, S \neq 0) = h(H_2 H_1^{-1}\mathbf{y}_1 + \mathbf{v}_2 - H_2 H_1^{-1}\mathbf{v}_1|\mathbf{y}_1, S \neq 0)$. But since when $\mathbf{y}_1$ is given, $H_2 H_1^{-1}\mathbf{y}_1 + \mathbf{v}_2 - H_2 H_1^{-1}\mathbf{v}_1$ is independent of the transmit power $\sigma^2$, we have $h(\mathbf{y}_2|\mathbf{y}_1, S \neq 0) = O(1)$ and so

$$h(\mathbf{y}|S \neq 0) = h(\mathbf{y}_1|S \neq 0) + O(1).$$

The entropy of the random vector $\mathbf{y}_1$ is upper bounded by the entropy of a zero-mean Gaussian random vector of the same variance, thus

$$h(\mathbf{y}|S \neq 0) \leq T \log \pi(1 + L\sigma^2) + O(1) = T \log \sigma^2 + O(1),$$

and so

$$h(\mathbf{y}) \leq (1-q)T \log \sigma^2 + O(1).$$

To find $h(\mathbf{y}|\mathbf{s})$, we note that $\mathbf{y}|_{\mathbf{s}}$ is a Gaussian random vector with covariance matrix $R_y = I_{T+L-1} + \sigma^2 SS^*$ and so

$$\begin{aligned}
h(\mathbf{y}|\mathbf{s}) &= E \log \det \pi \left(I_{T+L-1} + \sigma^2 SS^*\right) \\
&= \log \pi^{T+L-1} + E \log \det \left(I_L + \sigma^2 S^* S\right) \\
&= (1-q)E \log \det \left(I_L + \sigma^2 S^* S\right) + O(1) \\
&= (1-q) \log \det(\sigma^2 I_L) + (1-q)E \log \det \left(S^* S + \frac{1}{\sigma^2} I_L\right) + O(1) \\
&= (1-q)L \log \sigma^2 + O(1).
\end{aligned}$$

We therefore obtain

$$I(\mathbf{y};\mathbf{s}) \leq (1-q)(T-L) \log \sigma^2 + O(1).$$

The following upper bound is clearly maximized by choosing $q = 0$, i.e., by assuming that the transmitted matrix is generically nonsingular (non-zero). Thus, for any such distribution on $S$, we obtain $I(\mathbf{y};\mathbf{s}) \leq$

$(T-L)\log\sigma^2 + O(1)$, from which it follows that

$$C \leq (T-L)\log\sigma^2 + O(1). \tag{29}$$

On the other hand, Theorem 1 at high SNR yields $\rho_{\text{eff}} = \frac{\sigma^2 T}{(\sqrt{T-L}+\sqrt{L})^2}$. Thus, since $\bar{H}_d^*\bar{H}_d$ is generically nonsingular

$$\begin{aligned} C &\geq E\log\det\left(I + \frac{\sigma^2 T}{(\sqrt{T-L}+\sqrt{L})^2}\bar{H}_d^*\bar{H}_d\right) \\ &\geq E\log\det\left(\frac{\sigma^2 T}{(\sqrt{T-L}+\sqrt{L})^2}\bar{H}_d^*\bar{H}_d\right) \\ &= \log\det\sigma^2 I_{T-L} + O(1) \\ &= (T-L)\log\sigma^2 + O(1), \end{aligned} \tag{30}$$

where in the progression from the first to the second line we have used a well-known fact that for $A \geq 0$, $\det(I+A) \geq \det A$.

The inequalities (29) and (30) yield the desired result.

□

**Remark:** The above result shows that training-based schemes achieve the leading order term in the capacity at high SNR (a similar result for narrow-band flat-fading channels is given in [2, 7]). We may therefore claim that training-based schemes are optimal at high SNR. We should further mention that, although the results derived in this paper assumed that the channel coefficients are iid $\mathcal{CN}(0,1)$, this assumption can be considerably relaxed for the high SNR results. In fact, all that is needed is that

$$E\log\det\sigma^2\bar{H}^*\bar{H} = (T-L)\log\sigma^2 + O(1).$$

**Low SNR**

At low SNR, the issue is somewhat more tricky. It is wellknown (see, e.g., [1] and the references therein) that at low SNR, we have $C = O(\sigma^2)$. Examination of Theorem 1 on the other hand yields

$$C_\tau \geq O(\sigma^4). \tag{31}$$

At first glance it may appear that the above lower bound on the training-based capacity is tight since the effective noise term $\mathbf{v}'_d = \sigma_d \tilde{H}_d \mathbf{s}_d + \sigma_\tau \tilde{H}_\tau \theta_\tau + \mathbf{v}_d$ in (4) approaches Gaussian noise as $\sigma \to 0$, and so our lower bound obtained by replacing the effective noise by independent Gaussian noise should be tight. However, this claim cannot be true since in any scheme that uses training symbols, one can always ignore the fact that training was employed and for the remaining $T_d = T - T_\tau$ channel uses assume that the channel is unknown and obtain the (now scaled) low SNR bound $C = \frac{T_d}{T} O(\sigma^2) = O(\sigma^2)$. Of course, this is not what training schemes do: they assume that the channel estimate is perfect and ignore the correlation between the signal and effective additive noise.

Therefore, insofar as the *capacity* of any scheme that transmits training symbols goes, the bound in (31) is extremely loose at low SNR. However, if one considers the fact that training-based detection schemes assume perfect channel estimates and ignore the correlation between the signal and effective noise, then one realizes that training-based detectors are mismatched to the channel and that (31) is an indication of the mutual information obtainable from any such mismatch scheme. We therefore can claim that, at low SNR, training-based schemes are highly suboptimal since they can only deliver $O(\sigma^4)$ mutual information, as opposed to the true channel capacity $O(\sigma^2)$.[†]

# 6 Plots of Capacities

Figure 1 shows the training-based lower bounds on capacity as a function of the block length $T$ for $\sigma^2 = 6dB$ and the channel length $L = 4$. By allowing the training and data transmission powers to vary, we

[†]We should remark that the above claim is not entirely rigorous and merits a much more careful analysis. But this is a much more difficult problem to address as the problem of determining the mutual information obtainable by a mismatched decoder, as opposed to the true maximum-likelihood decoder, say, is an open problem.

achieve approximately $5 - 10\%$ increase in capacity. At $T = 50$, achieved capacity is approximately $20\%$ below the (unrealistic) capacity achieved when the receiver knows the channel perfectly.

In Figure 2, the optimal transmit power allocation $\sigma_d^2$ and $\sigma_\tau^2$ is plotted as a function of the block length. The dashed line in Figure 2 denotes the case of equal training and data transmission powers $\sigma_d^2 = \sigma_\tau^2$. Figure 2 illustrates what is implied by Theorem 1 — we need to spend more power for training than for transmission when $T > 2L$, more power for transmission than for training when $T < 2L$, and the same power for both when $T = 2L$.

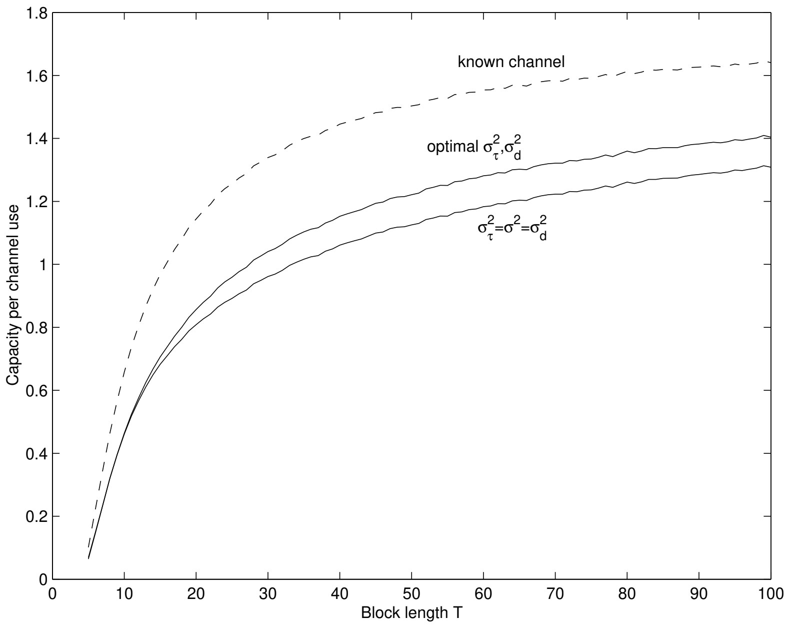


Figure 1: $C/T_d$ *vs.* $T$

# References


[1] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Transactions on Information Theory* , vol. 44, no. 6, Oct. 1998.

[2] B. Hassibi and B. M. Hochwald. "How Much Training is Needed in Multiple-Antenna Wireless Links?" submitted to *IEEE Transactions on Information Theory*.

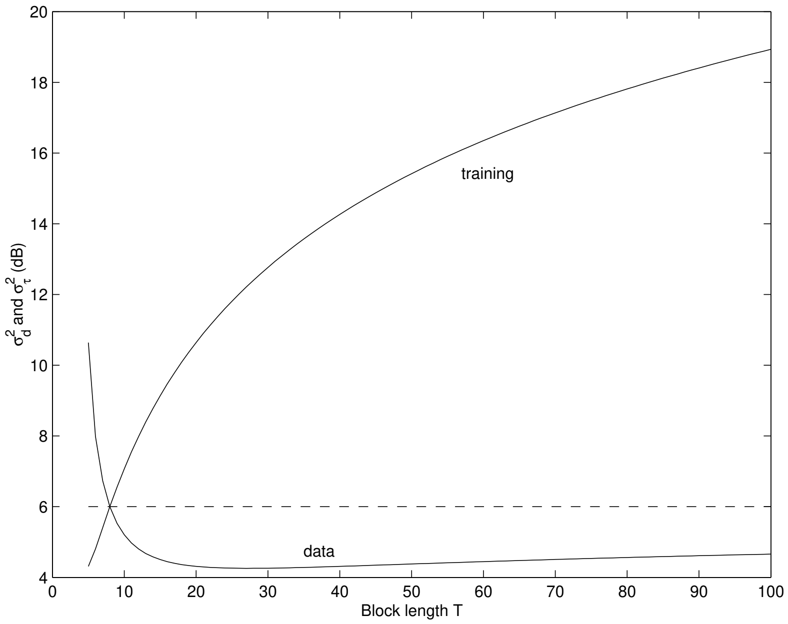


Figure 2: $\sigma_d^2, \sigma_\tau^2$ *vs.* $T$

[3] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel,: *IEEE Transactions on Information Theory*, vol. 46, no. 3, May 2000.

[4] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecom.*, vol. 10, pp. 585-595, Nov. 1999.

[5] L. Brandenburg and A. Wyner, "Capacity of the gaussian channel with memory: the multivariate case," *Bell Labs. Tech. J.*, vol. 53, pp. 745–778, 1974.

[6] T. N. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[7] 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.

[8] H. Vikalo, B. Hassibi, B. Hochwald, and T. Kailath, "Optimal training for frequency-selective channels," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 2105-2108.

[9] S. Adireddy, L. Tong, and H. Viswanathan, "Optimal placement of training for frequency-selective block-fading channels," *IEEE Transactions on Information Theory*, vol. 48, no. 8, August 2002.