Optimal Placement of Training for Frequency-Selective Block-Fading Channels

Srihari Adireddy, *Student Member, IEEE*, Lang Tong, *Senior Member, IEEE*, and Harish Viswanathan, *Member, IEEE*

***Abstract*—The problem of placing training symbols optimally for orthogonal frequency-division multiplexing (OFDM) and single-carrier systems is considered. The channel is assumed to be quasi-static with a finite impulse response of length $(L+1)$ samples. Under the assumptions that neither the transmitter nor the receiver knows the channel, and that the receiver forms a minimum mean square error (MMSE) channel estimate based on training symbols only, training is optimized by maximizing a tight lower bound on the ergodic training-based independent and identically distributed (i.i.d.) capacity. For OFDM systems, it is shown that the lower bound is maximized by placing the known symbols periodically in frequency. For single-carrier systems, under the assumption that the training symbols are placed in clusters of length $\alpha \geq (2L+1)$, it is shown that the lower bound is maximized by a family of placement schemes called QPP-α, where QPP stands for quasi-periodic placement. These placement schemes are formed by grouping the known symbols into as many clusters as possible and then placing these clusters periodically in the packet. For both OFDM and single-carrier systems, the optimum energy tradeoff between training and data is also obtained.**

***Index Terms*—Ergodic capacity, orthogonal frequency-division multiplexing (OFDM), placement schemes, single-carrier systems, training symbols, unknown channels.**

I. Introduction

THE problem of achieving the capacity of a linear, time-invariant Gaussian channel under the assumption that both the transmitter and the receiver know the channel is mature ([8] and the references in it). For wireless communications, especially mobile wireless, the channel is random and time-varying. Hence, the assumption that either the receiver or the transmitter knows the channel is unrealistic [3]. The rapid growth in mobile wireless applications has motivated the problem of finding the capacity of a fading channel under the assumption that neither the receiver nor the transmitter knows the channel (unknown channel scenario).

The block-fading model [3] provides a first-order approximation to the continuously time-varying channel, and it is simple enough to be mathematically tractable. The key parameter in this model is the coherence interval T. The channel is assumed to stay constant for T samples and change to a new value. The capacity of a single antenna system for the unknown channel scenario where the channel under goes Rayleigh flat-fading channel with $T=1$ has been addressed in [4]. The problem of finding the capacity for Rayleigh flat-fading model under a more general setting of multiple antennas and a general T was considered by Marzetta and Hochwald [15]. Their work gives useful insights for the single antenna problem as well. It was shown that as $T \rightarrow \infty$, the unknown channel capacity approaches the known channel capacity.

It is important to develop simple techniques that achieve the capacity of the unknown channel. A paradigm that is often employed in practice is to first estimate the unknown channel and then use the estimate to perform decoding. The most popular and practical technique of learning the channel is by insertion of training symbols in the data stream. While insertion of known symbols can in general be suboptimal, it is mandatory in order to simplify the receiver implementation. This introduces the notion of training-based capacity, which is the maximum rate achievable with codewords that consist of known and unknown symbols. The question then arises about how close the training-based capacity is to the capacity of the unknown channel and how one should optimize training to maximize the training-based capacity of a mobile wireless channel. This problem was considered for a multiple-antenna system under Rayleigh block fading scenario by Hassibi and Hochwald [2]. They obtained tight lower bounds on the capacity of the training-based systems and optimized the fraction of training symbols, energy allocated to training and data to maximize this bound. Their paper provides a useful framework for analyzing the capacity achievable by training-based schemes in general. An important insight of this analysis is that training is optimal at high signal-to-noise ratio (SNR) and suboptimal at low SNR. Similar techniques for lower-bounding mutual information under imperfect knowledge of the channel have been proposed by Medard [9].

Demand for higher bit rate leads to frequency-selective fading in mobile wireless channels. This motivates the question of designing training for frequency-selective fading channels with block fading. A new degree of freedom that is specific to frequency-selective channels is the placement of training. The performance for the flat fading scenario turns out to be independent of the placement of known symbols. Furthermore, the problem of training-symbol placement has to be addressed for both single-carrier and multicarrier systems separately since the paradigm for training is different for the two transmission systems.

Manuscript received June 21, 2001; revised February 22, 2002. This work was supported in part by the National Science Foundation under Contract CCR-9804019, Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564 and the ARL CTA on Communication and Networks. This work was performed in part when S. Adireddy was visiting Lucent Technologies, Murray Hill, NJ.

S. Adireddy and L. Tong are with the School of Electrical Engineering, 384 Frank H.T. Rhodes Hall, Cornell University, Ithaca, NY 14853 USA (e-mail: shrihari@ece.cornell.edu; ltong@ee.cornell.edu).

H. Viswanathan is with Lucent Technologies Bell Labs, Murray Hill, NJ 07974 USA (e-mail: harishv@lucent.com).

Communicated by G. Caire, Associate Editor for Communications.

Publisher Item Identifier 10.1109/TIT.2002.800466.

For single-carrier systems, the design of training, namely, the fraction of training, the choice of training symbols, and energy tradeoff between training and data, for frequency selective fading model was addressed in [7] under the assumption that all the training symbols are placed at the start of the packet. It was shown that at high SNR training-based schemes are capable of capturing most of the channel capacity, whereas at low SNR they are highly suboptimal. The placement of training though was assumed to be fixed.

The placement of training affects the capacity of the system through channel estimation and detection. We have previously considered the problem of joint optimization of symbol placement and equalizer for a symbol-by-symbol decision feedback receiver [11] under the assumption that the channel is known at the receiver. The performance criterion used was average mean-square error (AMSE). It turns out that the optimal symbol placement is to separate the known symbols by at least the detection delay d of the decision feedback receiver. The optimal placement of known symbols for single-carrier broadcast systems where the channel undergoes nonergodic fading was considered in [12]. The metric used was outage probability. It was shown that the outage probability is minimized by breaking the known symbols into small blocks and placing them periodically.

The problem of optimizing placement of training for minimizing the mean-square error (MSE) in channel estimate has been addressed for OFDM systems in [10]. Optimal training placement schemes were obtained for the more general setting of block precoded transmissions with cyclic prefix in [14]. The metric for optimization was again the MSE of the channel estimate. However, as alluded to earlier, channel estimation is just one facet of the problem. The placement of known symbols affects not only the channel estimate but also the detection of unknown symbols. In this paper, we take the holistic view and try to optimize the placement of known symbols by maximizing the training-based capacity.

In this paper, we first use the framework developed in [2] to obtain a tight lower bound on the training-based capacity of OFDM and single-carrier systems. We then optimize the placement of training by maximizing this lower bound. For OFDM systems, under the assumption that all the training symbols have equal energy, we show that the lower bound is maximized by placing the training symbols periodically in the OFDM symbol. That is, we pick equally spaced tones for training. This is the placement scheme that was also obtained in [10], [14]. It is remarkable that this placement not only gives the best channel estimate but also maximizes the tight lower bound on mutual information. For single-carrier systems, under the assumption that the training symbols are of length at least $\alpha \geq (2L+1)$, we show that the placement schemes in the class QPP-α (QPP stands for quasi-periodic placement) [12] are optimal. The placement schemes in QPP-α are obtained by breaking the known symbols into as many clusters as possible and placing them such that the unknown symbols blocks are as "equal" as possible.

This paper is organized as follows. In Section II, we introduce the system model. In Section III, we first formulate the optimization problem for OFDM systems and then determine optimal placement schemes. We consider the optimization of training for single-carrier systems in Section IV. In Section V, we illustrate the ideas through simulations, and finally, conclude in Section VI. The Appendix contains the proofs of lemmas and theorems stated in the paper.

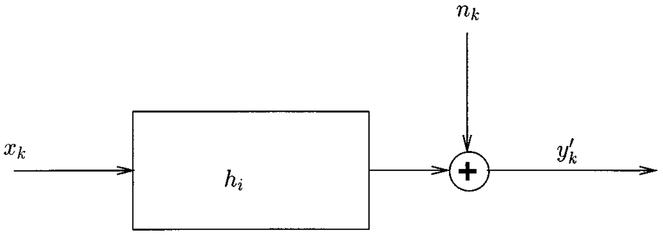

Fig. 1. System model.

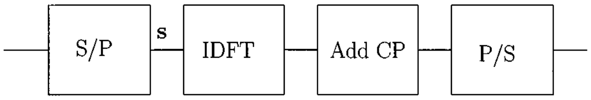

Fig. 2. Transmitter side processing.

II. System Model

The system model is shown in Fig. 1. The channel $\boldsymbol{h} = [h_0, h_1, \ldots h_L]^T$ has a finite-impulse response of length $(L+1)$ samples (where the symbol ^{T} denotes the transpose of the vector). We assume that taps of the channel $\boldsymbol{h}$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian with zero mean and variance equal to $\frac{1}{L+1}$. The fading coefficients remain constant for T symbol periods and change to an independent value. We assume that neither the receiver nor the transmitter knows the fading coefficients. The received signal is corrupted by additive white noise that is circularly symmetric complex Gaussian with zero mean and variance σ_w^2. This model, described above, is an extension of the quasi-static flat fading to quasi-static frequency-selective fading.

III. Optimal Placement Scheme and Training for OFDM System

A. OFDM System

Orthogonal frequency-division multiplexing (OFDM) has emerged as an attractive modulation scheme for high-data-rate communication systems. It is presently being used in standards like Digital Video Broadcast (DVB) and Digital Audio Broadcast (DAB). Proposals for fourth-generation systems include those that use OFDM as the modulation scheme. Fig. 2 shows the processing performed at the transmitter of the OFDM system. The symbol stream is parsed into blocks of length $(T-L)$ by the serial-to-parallel (S/P) converters. These blocks, called OFDM blocks, are then transformed by inverse discrete Fourier transform (IDFT). The cyclic prefix (CP) of length L is appended to each OFDM block to form a super block. We then perform a parallel-to-serial (P/S) conversion of the super blocks and transmit them. We assume that the channel stays constant over the duration of a super block.

Known symbols are introduced in frequency as is the norm for most OFDM standards. We assume that each OFDM block is of length $(N+P)$ where N is the number of unknown symbols and

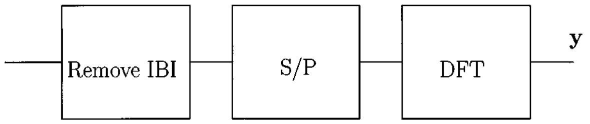

Fig. 3. Receiver side processing.

P is the number of known symbols (N and P are chosen such that $N+P=T-L$). The vector $\boldsymbol{s}=[s_1, s_2, \ldots, s_{(N+P)}]^T$ is formed by collecting the symbols in each OFDM block.

Fig. 3 illustrates the processing performed at the receiver. At the receiver, interblock interference (IBI), the output due to symbols from two different OFDM blocks, is dropped. The remaining data are parsed into blocks of length $(T-L)$ by the S/P converter and passed through the discrete Fourier transform (DFT). The vector $\boldsymbol{y}$ is formed by collecting the output corresponding to the block $\boldsymbol{s}$.

The channel is completely specified by the relation between the input $\boldsymbol{s}$ and the output $\boldsymbol{y}$. The channel law is given by

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N+P} \end{bmatrix} = \underbrace{\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & 0 \\ & & \ddots & \\ 0 & & & d_{N+P} \end{bmatrix}}_{\boldsymbol{D}} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N+P} \end{bmatrix} + \boldsymbol{w} \quad (1)$$

where d_i is the ith Fourier coefficient of the $(N+P)$-point DFT of the channel $\boldsymbol{h}$. That is,

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N+P} \end{bmatrix} = \sqrt{(N+P)}\,\boldsymbol{W}_L\boldsymbol{h} \quad (2)$$

where $\boldsymbol{W}_L$ is the truncated unit norm DFT matrix of size $(N+P)\times(L+1)$, i.e.,

$$[\boldsymbol{W}_L]_{kl} = \frac{1}{\sqrt{N+P}} \exp\frac{-j2\pi(k-1)(l-1)}{N+P}. \quad (3)$$

Intuitively speaking, the OFDM transmission scheme converts frequency-selective fading in time to flat fading on each tone. The vector $\boldsymbol{w}$ is zero mean, circular, Gaussian with covariance equal to $\sigma_w^2\boldsymbol{I}$.

B. Problem Statement

We now formulate the problem of designing optimal training. Training symbols are introduced in $\boldsymbol{s}$ to estimate the channel $\boldsymbol{h}$. We define $\mathcal{P}$ as the set of indexes of the tones used for training and as set $\mathcal{P}_c$, the indexes of the tones used for transmitting data. The placement scheme is completely specified by the set $\mathcal{P}$. We denote as $\boldsymbol{s}_t = (s_{1t}, \ldots, s_{Pt})^T$ the vector of symbols used for training. We use the subscript $1t$ to represent the smallest element of the set $\mathcal{P}$, and so on. Let $\boldsymbol{s}_d$ be the vector of data symbols, namely, $\boldsymbol{s}_d = (s_{1d}, \ldots, s_{Nd})^T$.

The power constraint on the system is formulated as

$$\frac{1}{(N+P)}\left(\mathrm{E}\{\operatorname{tr}\boldsymbol{s}_d\boldsymbol{s}_d^H\} + \operatorname{tr}\boldsymbol{s}_t\boldsymbol{s}_t^H\right) = 1. \quad (4)$$

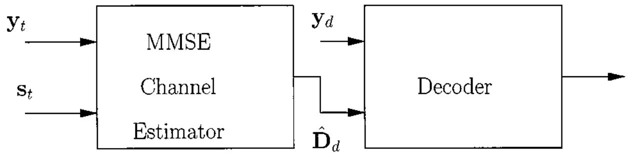

Fig. 4. Receiver structure.

We do not constrain the data and training powers to be the same. If $\rho_d = \frac{1}{N}\mathrm{E}\{\operatorname{tr}\boldsymbol{s}_d\boldsymbol{s}_d^H\}$ and $\rho_t = \frac{1}{P}\operatorname{tr}\boldsymbol{s}_t\boldsymbol{s}_t^H$, then (4) can be written as

$$\frac{N\rho_d + P\rho_t}{N+P} = 1. \quad (5)$$

We restrict ourselves to receivers of the structure given in Fig. 4. We define as $\boldsymbol{y}_t$ the output that is due to training. It is given by

$$\boldsymbol{y}_t = \underbrace{\begin{bmatrix} d_{1t} & & \\ & \ddots & \\ & & d_{Pt} \end{bmatrix}}_{\boldsymbol{D}_t} \begin{bmatrix} s_{1t} \\ \vdots \\ s_{Pt} \end{bmatrix} + \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_{Pt} \end{bmatrix}}_{\boldsymbol{w}_t}. \quad (6)$$

Similarly, $\boldsymbol{y}_d$ is defined as the output due to the data symbols. It is given by

$$\boldsymbol{y}_d = \boldsymbol{D}_d\boldsymbol{s}_d + \boldsymbol{w}_d \quad (7)$$

where $\boldsymbol{D}_d = \operatorname{diag}(d_{1d}, d_{2d}, \ldots, d_{Nd})$. We assume that the channel estimator forms the minimum mean-square error (MMSE) estimate of the channel using only training. The decoder then uses $\boldsymbol{y}_d$ and the MMSE estimate $\hat{\boldsymbol{D}}_d$ to perform the decoding. There is no loss in the restriction to linear MMSE estimators. This is due to the fact that for a channel with Gaussian statistics, we have

$$I(\boldsymbol{y}_d, \boldsymbol{y}_t; \boldsymbol{s}_d) = I(\boldsymbol{y}_d, \hat{\boldsymbol{D}}_d; \boldsymbol{s}_d). \quad (8)$$

This follows from the fact that the input distribution is independent of $\boldsymbol{y}_t$ and that $\boldsymbol{y}_d$ is independent of $\boldsymbol{y}_t$ given $\boldsymbol{s}_d$ and $\hat{\boldsymbol{D}}_d$.

We assume that the receiver performs optimal decoding, that is, in contrast to [1], the receiver does not assume that the channel estimate is perfect. The i.i.d. training-based capacity of the system is then equal to

$$C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) \triangleq \max_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} I(\boldsymbol{y}_d, \hat{\boldsymbol{D}}_d; \boldsymbol{s}_d) \quad (9)$$

where the probability distribution $f_{\text{i.i.d.}}(\boldsymbol{s}_d)$ and the training $\boldsymbol{s}_t$ are such that the input power constraint is satisfied. The notion of i.i.d. capacity used here is similar to the one in [13]. We also note that in this paper, by i.i.d. capacity, we in fact mean the i.i.d. training-based capacity.

Our objective then is to obtain optimal placement scheme $\mathcal{P}^*$, optimal energy allocation (ρ_d^*, ρ_t^*), and optimal training symbols $\boldsymbol{s}_t^*$ as

$$(\mathcal{P}^*, \rho_d^*, \rho_t^*, \boldsymbol{s}_t^*) = \arg\max_{\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t} C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t). \quad (10)$$

C. MMSE Channel Estimate

In this subsection, we obtain expressions for the MMSE estimate of the channel. The model for channel estimation is given by

$$\boldsymbol{y}_t = \underbrace{\begin{bmatrix} d_{1t} & & \\ & \ddots & \\ & & d_{Pt} \end{bmatrix}}_{\boldsymbol{D}_t} \underbrace{\begin{bmatrix} s_{1t} \\ \vdots \\ s_{Pt} \end{bmatrix}}_{\boldsymbol{s}_t} + \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_{Pt} \end{bmatrix}}_{\boldsymbol{w}_t}. \tag{11}$$

Equation (11) can be rewritten as

$$\boldsymbol{y}_t = \sqrt{N+P}\,\boldsymbol{S}\,\underbrace{\boldsymbol{I}_{\mathcal{P}}\boldsymbol{W}_L}_{\boldsymbol{W}_{\mathcal{P}L}}\,\boldsymbol{h} + \boldsymbol{w}_t \tag{12}$$

where the matrix $\boldsymbol{S}$ is given by $\operatorname{diag}(s_{1t}, \ldots, s_{Pt})$, the matrix $\boldsymbol{I}_{\mathcal{P}}$ is a selection matrix of size $P \times (N+P)$ with a 1 in row i at the position given the ith index in $\mathcal{P}$ and with 0's elsewhere. Using the fact that $\mathrm{E}\,\boldsymbol{h}\boldsymbol{h}^H = \frac{1}{L+1}\boldsymbol{I}$, we can write the MMSE estimate as

$$\begin{aligned}\hat{\boldsymbol{h}} = {} & \sqrt{N+P}\,\boldsymbol{W}_{\mathcal{P}L}^H\boldsymbol{S}^H \\ & \cdot \left((N+P)\boldsymbol{S}\boldsymbol{W}_{\mathcal{P}L}\boldsymbol{W}_{\mathcal{P}L}^H\boldsymbol{S}^H + (L+1)\sigma_w^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}_t.\end{aligned}$$

We also note that the covariance matrix of the error $\tilde{\boldsymbol{h}} = \boldsymbol{h} - \hat{\boldsymbol{h}}$ is given by

$$\mathrm{E}\{\tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H\} = \frac{1}{L+1}\left(\boldsymbol{I} + \frac{N+P}{\sigma^2}\boldsymbol{W}_{\mathcal{P}L}^H\boldsymbol{S}^H\boldsymbol{S}\boldsymbol{W}_{\mathcal{P}L}\right)^{-1} \tag{13}$$

where $\sigma^2 = (L+1)\sigma_w^2$. The covariance matrix of the estimate $\hat{\boldsymbol{h}}$ is given by

$$\mathrm{E}\,\hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^H = \frac{\boldsymbol{I}}{L+1} - \mathrm{E}\,\tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H. \tag{14}$$

From the estimate of $\boldsymbol{h}$, we can obtain $\hat{\boldsymbol{D}}_d$. If $\boldsymbol{d}_d = \operatorname{diag}(\boldsymbol{D}_d)$ is the vector formed by collecting the diagonal elements of $\boldsymbol{D}_d$ then $\hat{\boldsymbol{d}}_d$, the MMSE estimate of $\boldsymbol{d}_d$, can be written as

$$\hat{\boldsymbol{d}}_d = \sqrt{(N+P)}\,\boldsymbol{I}_{\mathcal{P}_c}\boldsymbol{W}_L\hat{\boldsymbol{h}} \tag{15}$$

where $\boldsymbol{I}_{\mathcal{P}_c}$ is a selection matrix of size $N \times (N+P)$ matrix with a 1 in row i at the position given by the ith index in $\mathcal{P}_c$ and with 0's elsewhere. The covariance of the error in the estimate of the data tones $\tilde{\boldsymbol{d}}_d = \boldsymbol{d}_d - \hat{\boldsymbol{d}}_d$ is given by

$$\begin{aligned}\mathrm{E}\{\tilde{\boldsymbol{d}}_d\tilde{\boldsymbol{d}}_d^H\} = {} & \sqrt{\frac{N+P}{L+1}}\,\boldsymbol{I}_{\mathcal{P}_c}\boldsymbol{W}_L \\ & \cdot \left(\boldsymbol{I} + \frac{N+P}{\sigma^2}\boldsymbol{W}_{\mathcal{P}L}^H\boldsymbol{S}^H\boldsymbol{S}\boldsymbol{W}_{\mathcal{P}L}\right)^{-1}\boldsymbol{W}_L^H\boldsymbol{I}_{\mathcal{P}_c}^H\sqrt{\frac{N+P}{L+1}}.\end{aligned} \tag{16}$$

D. Lower Bound on Training-Based Capacity

In this section we obtain a tight lower bound for $C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t)$ and optimize training with respect to this bound. We have

$$\begin{aligned}C(\mathcal{P}, \boldsymbol{s}_t, \rho_d, \rho_t) &= \max_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} I(\boldsymbol{y}_d; \boldsymbol{s}_d|\hat{\boldsymbol{D}}) + \underbrace{I(\hat{\boldsymbol{D}}; \boldsymbol{s}_d)}_{=0} \\ &= \max_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} I(\boldsymbol{y}_d; \boldsymbol{s}_d|\hat{\boldsymbol{D}})\end{aligned} \tag{17}$$

because the MMSE estimate $\hat{\boldsymbol{D}}$ is independent of $\boldsymbol{s}_d$. The relationship between $\boldsymbol{y}_d$ and $\boldsymbol{s}_d$ is given by

$$\boldsymbol{y}_d = \underbrace{\begin{bmatrix} d_{1d} & & \\ & \ddots & \\ & & d_{Nd} \end{bmatrix}}_{\boldsymbol{D}_d} \begin{bmatrix} s_{1d} \\ \vdots \\ s_{Nd} \end{bmatrix} + \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_{Nd} \end{bmatrix}}_{\boldsymbol{w}_d}. \tag{18}$$

This can be rewritten as

$$\boldsymbol{y}_d = \hat{\boldsymbol{D}}_d\boldsymbol{s}_d + \underbrace{\tilde{\boldsymbol{D}}_d\boldsymbol{s}_d + \boldsymbol{w}_d}_{\boldsymbol{\nu}_d} \tag{19}$$

where $\hat{\boldsymbol{D}}_d$ is the estimate of $\boldsymbol{D}_d$ and $\tilde{\boldsymbol{D}}_d$ the error in the estimate. It is difficult to evaluate the i.i.d. capacity because the distribution of $\boldsymbol{\nu}_d$ is difficult to characterize. Therefore, we obtain a lower bound on the i.i.d. channel capacity and then reformulate the problem of optimization in terms of this lower bound.

The lower bound is obtained as follows. Given $f_{\text{i.i.d.}}(\boldsymbol{s}_d)$, we define $\Omega(f_{\text{i.i.d.}}(\boldsymbol{s}_d))$, the set of all the conditional probability distributions for a random variable that has the same first- and second-order properties as $\boldsymbol{\nu}_d$. That is,

$$\begin{aligned}&\Omega(f_{\text{i.i.d.}}(\boldsymbol{s}_d)) = \{p(\boldsymbol{n}_d/\boldsymbol{s}_d, \hat{\boldsymbol{D}})\colon \mathrm{E}\{\boldsymbol{n}_d/\hat{\boldsymbol{D}}\} = \mathrm{E}\{\boldsymbol{\nu}_d/\hat{\boldsymbol{D}}\} \\ &\mathrm{E}\{\boldsymbol{n}_d\boldsymbol{n}_d^H/\hat{\boldsymbol{D}}\} = \mathrm{E}\{\boldsymbol{\nu}_d\boldsymbol{\nu}_d^H/\hat{\boldsymbol{D}}\}, \mathrm{E}\{\boldsymbol{n}_d\boldsymbol{s}_d^H/\hat{\boldsymbol{D}}\} = \mathrm{E}\{\boldsymbol{\nu}_d\boldsymbol{s}_d^H/\hat{\boldsymbol{D}}\}.\end{aligned} \tag{20}$$

Due to the properties of the MMSE estimator for Gaussian channels, we have

$$\begin{aligned}\mathrm{E}\{\boldsymbol{\nu}_d/\hat{\boldsymbol{D}}\} &= 0 \\ \mathrm{E}\{\boldsymbol{\nu}_d\boldsymbol{s}_d^H/\hat{\boldsymbol{D}}\} &= 0.\end{aligned}$$

Now consider the new model

$$\boldsymbol{z}_d = \hat{\boldsymbol{D}}_d\boldsymbol{s}_d + \boldsymbol{n}_d. \tag{21}$$

For this model, we consider the following quantity:

$$\begin{aligned}&C_{\text{lb}}(\mathcal{P}, \boldsymbol{s}_t, \rho_d, \rho_t) \\ &\quad \triangleq \sup_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} \inf_{p(\boldsymbol{n}_d/\boldsymbol{s}_d, \hat{\boldsymbol{D}}_d)} I(\boldsymbol{z}_d, \hat{\boldsymbol{D}}_d; \boldsymbol{s}_d).\end{aligned} \tag{22}$$

It is easy to see that $C_{\text{lb}}(\cdot)$ is a lower bound on $C(\cdot)$. This method of lower bounding is similar to the one used in [2].

Theorem 1: We have

$$C_{\text{lb}}(\mathcal{P}, \boldsymbol{s}_t, \rho_d, \rho_t) = \mathrm{E}\,\log\det\left(\boldsymbol{I} + \rho_d\boldsymbol{R}_{\boldsymbol{\nu}}^{-1}\hat{\boldsymbol{D}}_d\hat{\boldsymbol{D}}_d^H\right) \tag{23}$$

where $\boldsymbol{R}_{\boldsymbol{\nu}}$ is the autocorrelation of $\boldsymbol{\nu}_d$ and the expectation is with respect to the random variables in $\hat{\boldsymbol{D}}_d$.

Proof: Please refer to Appendix I.

Therefore, we have

$$\begin{aligned}C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) &\geq C_{\text{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) \\ &= \mathrm{E}\,\log\det\left(\boldsymbol{I} + \rho_d\boldsymbol{R}_{\boldsymbol{\nu}}^{-1}\hat{\boldsymbol{D}}_d\hat{\boldsymbol{D}}_d^H\right).\end{aligned} \tag{24}$$

At low SNR, $\boldsymbol{\nu}_d$ is close to Gaussian and the bound is tight. We conjecture, that using the same arguments as in [2], [7], the bound is tight at high SNR.

The conditional autocorrelation of $\boldsymbol{\nu}_d$ is given by

$$\begin{aligned} \boldsymbol{R}_{\boldsymbol{\nu}} &= \mathrm{E}\{\boldsymbol{\nu}_d \boldsymbol{\nu}_d^H / \boldsymbol{D}_d\} \\ &= \mathrm{E}\left\{\tilde{\boldsymbol{D}}_d \boldsymbol{s}_d \boldsymbol{s}_d^H \tilde{\boldsymbol{D}}_d^H\right\} + \mathrm{E}\{\boldsymbol{w}_d \boldsymbol{w}_d^H\} \\ &= \rho_d \boldsymbol{R}_e + \sigma_w^2 \boldsymbol{I}. \end{aligned}$$

The matrix $\boldsymbol{R}_e$ is diagonal since the symbols s_{kd} and s_{ld} are independent for $k \neq l$. The ith diagonal entry in $\boldsymbol{R}_e$ denoted as k_i is the MSE of the ith data tone. It can be obtained as

$$k_i = \boldsymbol{p}_i \mathrm{E}\left\{\tilde{\boldsymbol{d}}_d \tilde{\boldsymbol{d}}_d^H\right\} \boldsymbol{p}_i^H \tag{25}$$

where $\boldsymbol{p}_i$ is a row vector with a 1 in the index of the ith data tone and 0's elsewhere. In terms of the MSE of the data tones, $C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t)$ can be written as

$$C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) = \sum_{i=1}^{N} \mathrm{E}\left\{\log\left(1 + \frac{\rho_d \hat{d}_{id} \hat{d}_{id}^*}{\rho_d k_i + \sigma_w^2}\right)\right\} \tag{26}$$

where $\hat{d}_{id}$ is the estimate of the ith data tone. We normalize the Gaussian random variable $\hat{d}_{id}$ by dividing by the standard deviation and obtain the zero mean, unit variance Gaussian random variable $\bar{d}_{id}$. That is, $\hat{d}_{id} = \sqrt{1-k_i}\,\bar{d}_{id}$. The lower bound can be rewritten as

$$\begin{aligned} C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) &= \sum_{i=1}^{N} \mathrm{E}\,\log\left(1 + \frac{\rho_d(1-k_i)}{\rho_d k_i + \sigma_w^2}\,\bar{d}_{id}\bar{d}_{id}^*\right) \\ &= \sum_{i=1}^{N} f\left(\frac{1-k_i}{\gamma_d + k_i}\right) \end{aligned} \tag{27}$$

where $\gamma_d = \frac{\sigma_w^2}{\rho_d}$ is the inverse data SNR. The function $f(\cdot)$ is defined as

$$f(n) = \mathrm{E}\,\log\left(1 + n|x|^2\right) \tag{28}$$

where x is a complex Gaussian random variable with zero mean and unit variance. We observe that the capacity lower bound is a function of the MSE of the data tones alone and not those of the training tones.

E. Optimal Placement of Training

In this section, we optimize the placement by maximizing the lower bound on capacity. At the outset, we assume all the training symbols are constrained to be of equal energy, that is, $|s_{it}|^2 = \rho_t$, $i = 1, \ldots, P$. This is the case for most of the current OFDM systems. We, however, do not claim the optimality of equal energy allotment

$$C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) = C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, |\boldsymbol{s}_t|).$$

From (25) and (27), we note that the lower bound $C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t)$ depends only on the magnitude of the training symbols $\boldsymbol{s}_t$ and hence $C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t)$ is a function of only $\mathcal{P}$, ρ_d and ρ_t. For equal energy training schemes, we therefore exclude $\boldsymbol{s}_t$ as an argument of C_{lb}.

We obtain the optimal values of placement and energy tradeoff $(\mathcal{P}^*, \rho_d^*, \rho_t^*)$ as

$$\begin{aligned} (\mathcal{P}^*, \rho_d^*, \rho_t^*) &= \arg\max_{\mathcal{P}, \rho_d, \rho_t} C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t) \\ &= \arg\max_{\mathcal{P}, \rho_d, \rho_t} \sum_{i=1}^{N} f\left(\frac{1-k_i}{\gamma_d + k_i}\right). \end{aligned} \tag{29}$$

We attack the problem of joint optimization, by first fixing the energy tradeoff and maximizing the lower bound with respect to the placement. We first have the following lemma.

Lemma 1: For any given energy tradeoff (ρ_d, ρ_t), we have

$$C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t) \le N f\left(-1 + \frac{1+\gamma_d}{N} \max_{\mathcal{P}} \sum_{i=1}^{N} \frac{1}{\gamma_d + k_i}\right). \tag{30}$$

Proof: Refer to Appendix II.

Next, we maximize $\sum_{i=1}^{N} \frac{1}{\gamma_d + k_i}$ over the set of all possible placements and obtain an upper bound on $C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t)$, which is a function of only ρ_d and ρ_t.

Lemma 2: The lower bound satisfies

$$C_{\mathrm{lb}}(\mathcal{P}, \rho_d, \rho_t) \le N f\left(\frac{P}{\gamma_t(\gamma_d + 1) + P\gamma_d}\right) \tag{31}$$

where $\gamma_t = \frac{\sigma_w^2}{\rho_t}$ is the inverse training SNR.

Proof: Refer to Appendix III.

We now show that a simple placement scheme achieves this upper bound and is thus optimal for any energy tradeoff. Consider the placement $\mathcal{P}^*$ obtained by selecting the training tones periodically. We assume that N is a multiple of P so that such a selection is possible. It is easy to verify that if $P \ge (L+1)$, then for this placement, the matrix $\boldsymbol{W}_{\mathcal{P}L}^H \boldsymbol{W}_{\mathcal{P}L}$ is a multiple of the identity matrix. From (16) and (25), we find that

$$k_i = \frac{\gamma_t}{P + \gamma_t}, \qquad \forall i. \tag{32}$$

From (27) we have

$$C_{\mathrm{lb}}(\mathcal{P}^*, \rho_d, \rho_t) = N f\left(\frac{P}{\gamma_t(\gamma_d + 1) + P\gamma_d}\right) \tag{33}$$

and from Lemma 2 we conclude that $\mathcal{P}^*$ is optimal. We hence have the following theorem.

Theorem 2: For any energy tradeoff (ρ_d, ρ_t), under the assumption that $N = mP (m \ge 1)$, and $P \ge (L+1)$, all of the following placements are optimal:

$$\mathcal{P}^* = \{i,\, i+m+1,\, i+2(m+1),\, \ldots,\, i+(P-1)(m+1)\} \tag{34}$$

where i can take values from 1 to $(m+1)$. For any of these placements, the lower bound is given by

$$\begin{aligned} &C_{\mathrm{lb}}(\mathcal{P}^*, \rho_d, \rho_t) \\ &\quad = N\mathrm{E}\,\log\left(1 + \frac{\rho_d}{\sigma_w^2}\,\frac{P\rho_t dd^*}{P\rho_t + (L+1)(\rho_d + \sigma_w^2)}\right) \end{aligned} \tag{35}$$

where d is a complex Gaussian random variable with zero mean and unit variance.

It was shown in [10], [14] that the same set of placements minimizes the MSE in the estimate of $\boldsymbol{h}$. Their performance metric is hence $\sum_{i=1}^{N+P} k_i$, the sum of MSE of both data and training tones. Our performance metric is quite different. In fact, the capacity lower bound depends explicitly only on the MSE of data tones and not on those of training tones. To prove the optimality of periodic tone placement with respect to MSE, it is only necessary to show that this placement minimizes the arithmetic mean (AM) of the MSE of the data tones. But, in order to show optimality with respect to i.i.d. capacity, we show that the optimal placement minimizes the harmonic mean (HM) of the MSE of data tones. This is a stronger result than the previous one because for every placement scheme other than the optimal one, the HM of the MSEs is smaller than their AM. For the optimal placement, the HM is equal to the AM because the MSE for all the data tones is equal. It is, therefore, quite surprising that the same set of placements is optimal for this metric as well.

The obtained placement is optimal for any energy allocation. We assume that the training symbols are placed in optimal positions and optimize the energy allocation.

Theorem 3: Under the assumption that $N = mP$ $(m \geq 1)$ and $P \geq (L+1)$, the optimal energy distribution is given by

$$\begin{aligned} \rho_d^* &= \left(\sqrt{\gamma} - \sqrt{\gamma - 1}\right) \frac{\sqrt{\gamma}}{g} \\ \rho_t^* &= \left(\sqrt{\gamma} - \sqrt{\gamma - 1}\right) \frac{1}{\sqrt{hk}} \end{aligned} \tag{36}$$

where

$$\begin{aligned} h &= \frac{P}{(N+P)} \\ g &= \frac{N}{(N+P)} \\ k &= \frac{P(N-L-1)}{(L+1)((N+P)+N\sigma_w^2)} \end{aligned}$$

and

$$\gamma = \frac{h}{k} + 1.$$

Proof: Refer to Appendix IV.

The ratio of power in data to that in training is given by

$$\frac{g\rho_d^*}{h\rho_t^*} = \sqrt{1 + \frac{N-L-1}{(L+1)(1+g\sigma_w^2)}}. \tag{37}$$

At low SNR, we find that this ratio is equal to 1. Hence, half the energy is spent in training. Similar conclusions were reached in [2].

IV. Optimal Placement for Single-Carrier Systems

A. Single-Carrier System

Fig. 5 shows the processing performed at the transmitter of the single-carrier system. We assume that the symbols are parsed into packets of length $(T - L)$ by the S/P converters. A known symbol cluster $\boldsymbol{s}_k$ of length L is appended to the beginning of each block to form a super block. These known symbol clusters serve to remove the IBI between consecutive blocks and facilitate block-by-block processing. A P/S conversion is then performed on these superblocks and they are then transmitted through the channel. We have already mentioned that the channel stays constant for T samples and jumps to a new independent value (block-fading model). It is also necessary to specify over which part of the packet, the channel stays constant. As shown in Fig. 6, we assume that the channel stays constant from $t = 1$ to $t = T$. Over the period for which the channel stays constant we have

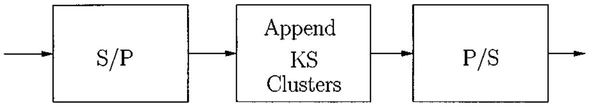

Fig. 5. Processing performed at the single-carrier transmitter.

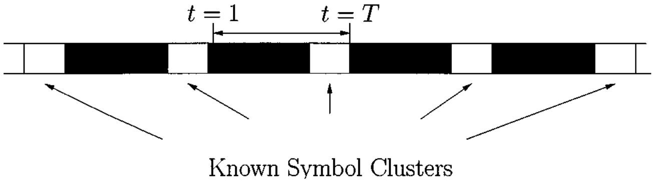

Fig. 6. The period over which the channel stays constant.

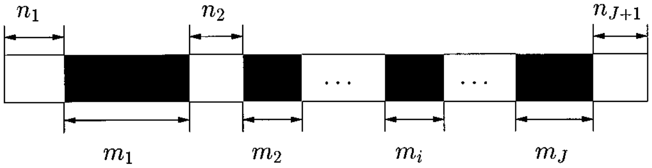

Fig. 7. Representation of placement schemes.

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_T \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} h_L & \cdots & h_0 & & & \\ & h_L & \cdots & h_0 & & \\ & & \ddots & & \ddots & \\ & & & \ddots & & \ddots \\ & & & h_L & \cdots & h_0 \end{bmatrix}}_{\boldsymbol{H}} \begin{bmatrix} \boldsymbol{s}_k \\ s_1 \\ \vdots \\ s_{N+P} \\ \boldsymbol{s}_k \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ \vdots \\ w_T \end{bmatrix} \tag{38}$$

where $\boldsymbol{h} = [h_0, \ldots, h_L]^T$ is a realization of the channel. We note that the output vector $\boldsymbol{y}$ is a function of both the symbols in the current packet $\boldsymbol{s} = [\boldsymbol{s}_k^T, s_1, \ldots, s_T]^T$ and the known symbol cluster $\boldsymbol{s}_k$ at the start of the next packet.

Each packet $\boldsymbol{s}$ consists of N unknown and $(P+L)$ known symbols. The known symbols are placed in clusters of length equal to $\alpha \geq L$. Fig. 7 shows the placement scheme of the vector $[\boldsymbol{s}^T \ \boldsymbol{s}_k^T]^T$. In general every placement can be specified by two tuples $(\boldsymbol{m}, \boldsymbol{n})$ where $\boldsymbol{m} = (m_1, \ldots, m_J)$ and $\boldsymbol{n} = (n_1, \ldots, n_{J+1})$. The tuple $\boldsymbol{m}$ gives the lengths of unknown

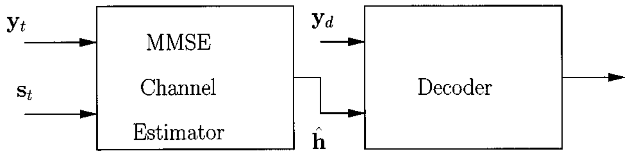

Fig. 8. Receiver structure.

symbol blocks and $\boldsymbol{n}$ gives the lengths of known symbol clusters. Since every packet starts with at least L known symbols, we know that n_1 is at least as big as L. We also note that the known symbol cluster $J+1$ includes the first L known symbols at the start of the next packet. Hence n_{J+1} is also at least as big as L. The minimum value of J is equal to 1, which corresponds to placing all the training at the ends of the packet. We note that the number of elements in each tuple is a function of the placement scheme. We refer to the symbols between any two consecutive known symbol clusters as unknown symbol blocks. Let the set $\mathcal{P}$ be the set of all possible placement schemes $(\boldsymbol{m}, \boldsymbol{n})$.

As shown in Fig. 8, the receiver consists of a channel estimator block followed by a decoder. The channel estimator forms an estimate of the channel based on training only. Since the channel varies from block to block, we can only form a block-by-block estimate of the channel. If s_{it}^k denotes the kth training symbol in the ith cluster, then we define the vector of training symbols

$$\boldsymbol{s}_t = [s_{1t}^1 \cdots s_{1t}^{n_1} \cdots s_{(J+1)t}^1 \cdots s_{(J+1)t}^{n_{J+1}}]^T.$$

We note again that

$$\boldsymbol{s}_k = [s_{1t}^1 \cdots s_{1t}^L]^T = [s_{(J+1)t}^{n_{J+1}-L+1} \cdots s_{(J+1)t}^{n_{J+1}}]^T.$$

We define as $\boldsymbol{y}_t$ the part of the output vector $\boldsymbol{y}$ that is due to training alone. The remaining part of the output vector is grouped as $\boldsymbol{y}_d$. The channel estimator block forms the estimate of the channel $\hat{\boldsymbol{h}} = g(\boldsymbol{y}_t, \boldsymbol{s}_t)$. As before, due the assumption that the channel is Gaussian, there is no loss in the restriction to linear MMSE estimators. The decoder uses $\boldsymbol{y}_d$, $\hat{\boldsymbol{h}}$, and $\boldsymbol{s}_t$ to perform the decoding.

We define as $\boldsymbol{s}_d$ the vector containing all the data symbols. The power constraint on the system is formulated as follows:

$$\frac{1}{(N+P+L)} \left(\mathrm{E}\{\mathrm{tr}\, \boldsymbol{s}_d \boldsymbol{s}_d^H\} + \mathrm{tr}\, \boldsymbol{s}_t \boldsymbol{s}_t^H\right) = 1. \tag{39}$$

We do not constrain the data and training powers to be the same. If $\rho_d = \frac{1}{N} \mathrm{E}\{\mathrm{tr}\, \boldsymbol{s}_d \boldsymbol{s}_d^H\}$ and $\rho_t = \frac{1}{P+L} \mathrm{tr}\, \boldsymbol{s}_t \boldsymbol{s}_t^H$, then (39) can be written as

$$\frac{N\rho_d + (P+L)\rho_t}{N+P+L} = 1. \tag{40}$$

B. Problem Statement

We now formulate the problem of optimal placement of training for single-carrier systems. The i.i.d. capacity of the system [13] can be defined as

$$C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) \triangleq \max_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} I(\boldsymbol{y}_d, \hat{\boldsymbol{h}}; \boldsymbol{s}_d) \tag{41}$$

where the probability distribution $f_{\text{i.i.d.}}(\boldsymbol{s}_d)$ and the training $\boldsymbol{s}_t$ are such that the input power constraint is satisfied. Our objective then is to obtain the optimal placement scheme $\mathcal{P}^*$, optimal energy tradeoff (ρ_d^*, ρ_t^*), and optimal training symbols $\boldsymbol{s}_t^*$ as

$$(\mathcal{P}^*, \rho_d^*, \rho_t^*, \boldsymbol{s}_t^*) = \arg \max_{\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t} C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t). \tag{42}$$

C. Training-Only Based MMSE Channel Estimate

In this subsection, we give properties of the channel estimator block. We assume that the estimator forms the MMSE estimate of the channel. The model for channel estimation is given by

$$\boldsymbol{y}_t = \boldsymbol{S}_t \boldsymbol{h} + \boldsymbol{w}_t \tag{43}$$

where

$$\boldsymbol{S}_t = \begin{bmatrix} \boldsymbol{S}_{1t} \\ \boldsymbol{S}_{2t} \\ \vdots \\ \boldsymbol{S}_{(J+1)t} \end{bmatrix}. \tag{44}$$

The matrix $\boldsymbol{S}_{it}$ is a Toeplitz matrix of size $(n_i - L) \times (L+1)$. It is formed by the training symbols in the ith training cluster as

$$\boldsymbol{S}_{it} = \begin{bmatrix} s_{it}^{(L+1)} & \cdots & s_{it}^1 \\ \vdots & \ddots & \\ s_{it}^{n_i} & \cdots & s_{it}^{(n_i-L)} \end{bmatrix}. \tag{45}$$

It is easy to see that the matrix $\boldsymbol{S}_t$ is of size

$$(P - (J-1)L) \times (L+1).$$

The MMSE estimate can then be written as

$$\hat{\boldsymbol{h}} = \boldsymbol{S}_t^H (\boldsymbol{S}_t \boldsymbol{S}_t^H + (L+1)\sigma_w^2 \boldsymbol{I})^{-1} \boldsymbol{y}_t. \tag{46}$$

The covariance of the error $\tilde{\boldsymbol{h}}$ is given by

$$\mathrm{E}\left\{\tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H\right\} = \frac{1}{(L+1)} \left(\boldsymbol{I} + \frac{\boldsymbol{S}_t^H \boldsymbol{S}_t}{\sigma^2}\right)^{-1} \tag{47}$$

where $\sigma^2 = (L+1)\sigma_w^2$. The covariance matrix of the estimate $\hat{\boldsymbol{h}}$ is given by

$$\mathrm{E}\left\{\hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^H\right\} = \frac{\boldsymbol{I}}{L+1} - \mathrm{E}\left\{\tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H\right\}. \tag{48}$$

We restrict ourselves to the case of orthogonal training that is the matrix $\boldsymbol{S}_t^H \boldsymbol{S}_t = c\boldsymbol{I}$ where c is a constant. This restriction is primarily motivated by simpler receiver implementation and mathematical tractability and we do not claim that this choice is optimal. The power constraint on training implies that

$$c \le (P+L)\rho_t. \tag{49}$$

Orthogonal training also imposes the upper bound on the number of clusters J. The matrix $\boldsymbol{S}_t$ has to be tall and hence

$$P - L(J-1) \ge (L+1).$$

This implies that

$$J \le \left\lfloor \frac{P-1}{L} \right\rfloor. \tag{50}$$

Further, $P \ge (L+1)$. The restriction to orthogonal training also implies that the taps of $\tilde{\boldsymbol{h}}$ are independent.

D. Lower Bound on Training-Based Capacity

In this subsection, since the problem of evaluating the i.i.d. capacity is complicated, we obtain a tight lower bound for $C(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t)$ and optimize training for this bound. As earlier, we have

$$C(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t) = \max_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} I(\boldsymbol{y}_d;\, \boldsymbol{s}_d|\hat{\boldsymbol{h}}). \tag{51}$$

The relationship between $\boldsymbol{y}_d$ and $\boldsymbol{s}_d$ is given by

$$\underbrace{\begin{bmatrix}\boldsymbol{y}_{1d}\\ \boldsymbol{y}_{2d}\\ \vdots\\ \boldsymbol{y}_{Jd}\end{bmatrix}}_{\boldsymbol{y}_d} = \underbrace{\begin{bmatrix}\boldsymbol{H}_{m_1} & 0 & \cdots & 0\\ 0 & \boldsymbol{H}_{m_2} & & 0\\ \vdots & & \ddots & \\ 0 & & & \boldsymbol{H}_{m_J}\end{bmatrix}}_{\boldsymbol{H}_d} \underbrace{\begin{bmatrix}\boldsymbol{s}_{1d}\\ \boldsymbol{s}_{2d}\\ \vdots\\ \boldsymbol{s}_{Jd}\end{bmatrix}}_{\boldsymbol{s}_d} + \underbrace{\begin{bmatrix}\boldsymbol{T}_1\\ \boldsymbol{T}_2\\ \vdots\\ \boldsymbol{T}_J\end{bmatrix}}_{\boldsymbol{T}} \boldsymbol{h} + \boldsymbol{w}_d. \tag{52}$$

The matrix $\boldsymbol{H}_{m_i}$ is a Toeplitz matrix of size $(m_i + L) \times m_i$ given by

$$\boldsymbol{H}_{m_i} = \begin{bmatrix} h_0 & 0 & \cdots & 0\\ h_1 & h_0 & & \vdots\\ \vdots & h_1 & & 0\\ h_L & \vdots & & h_0\\ 0 & h_L & & h_1\\ \vdots & 0 & & \vdots\\ 0 & \cdots & \cdots & h_L \end{bmatrix}_{(m_i+L)\times(m_i)}. \tag{53}$$

The fact that each training symbol cluster is at least as long as α where $\alpha \geq L$ leads to the matrix $\boldsymbol{H}_d$ being block-diagonal with $\boldsymbol{H}_{m_i}$ having the structure shown above. The matrix $\boldsymbol{H}_d$ is not block-diagonal if the training symbol clusters are allowed to be smaller than L. The vector $\boldsymbol{s}_{id}$ is of length m_i and is composed of data symbols in the ith unknown symbol block. The matrix $\boldsymbol{T}_i$ is composed of the training symbols $(s_{it}^{n_i-L+1},\, \ldots,\, s_{it}^{n_i})$ and $(s_{(i+1)t}^{1},\, \ldots,\, s_{(i+1)t}^{L})$. That is, $\boldsymbol{T}_i$ is a function of the L training symbols immediately before and after the ith unknown symbol block. These matrices are introduced to account for the fact that the first L and the last L samples of $\boldsymbol{y}_{id}$ are affected by the training symbols.

We can express $\boldsymbol{y}_d$ in terms of the estimate $\hat{\boldsymbol{h}}$ and the error $\tilde{\boldsymbol{h}}$ as

$$\boldsymbol{y}_d = \hat{\boldsymbol{H}}_d\boldsymbol{s}_d + \boldsymbol{T}\hat{\boldsymbol{h}} + \tilde{\boldsymbol{H}}_d\boldsymbol{s}_d + \boldsymbol{T}\tilde{\boldsymbol{h}} + \boldsymbol{w}_d. \tag{54}$$

We subtract $\boldsymbol{T}\hat{\boldsymbol{h}}$ from $\boldsymbol{y}_d$ to obtain $\boldsymbol{y}_d'$. We thus have

$$\boldsymbol{y}_d' = \hat{\boldsymbol{H}}_d\boldsymbol{s}_d + \underbrace{\tilde{\boldsymbol{H}}_d\boldsymbol{s}_d + \boldsymbol{T}\tilde{\boldsymbol{h}} + \boldsymbol{w}_d}_{\boldsymbol{\nu}_d}. \tag{55}$$

It is easy to see that

$$I(\boldsymbol{y}_d;\, \boldsymbol{s}_d|\hat{\boldsymbol{h}}) = I(\boldsymbol{y}_d';\, \boldsymbol{s}_d|\hat{\boldsymbol{h}}). \tag{56}$$

But it is difficult to obtain the latter analytically. As in Section III, we obtain a lower bound on the i.i.d. channel capacity by varying the conditional distribution of the noise among those that have the same first- and second-order properties as $\boldsymbol{\nu}_d$, namely, $\mathrm{E}\{\boldsymbol{\nu}_d/\hat{\boldsymbol{h}}\} = 0$, the conditional autocorrelation is given by

$$\mathrm{E}\left\{\boldsymbol{\nu}_d\boldsymbol{\nu}_d^H/\hat{\boldsymbol{h}}\right\} = \rho_d\mathrm{E}\,\tilde{\boldsymbol{H}}_d\tilde{\boldsymbol{H}}_d^H + \frac{1}{(L+1)\left(1+\frac{c}{\sigma^2}\right)}\boldsymbol{T}\boldsymbol{T}^H + \sigma_w^2\boldsymbol{I} \tag{57}$$

$$\triangleq \boldsymbol{R}_{\boldsymbol{\nu}}. \tag{58}$$

We also note that $\mathrm{E}\{\boldsymbol{\nu}_d\boldsymbol{s}_d^H/\hat{\boldsymbol{h}}\} = 0$ due to the property of the MMSE estimator.

We obtain a lower bound on the training-based capacity by an argument similar to one in Theorem 1. It can be shown that the worst case noise is zero mean Gaussian with autocorrelation $\boldsymbol{R}_{\boldsymbol{\nu}}$ and is independent of $\boldsymbol{s}_d$. Therefore, we have

$$\begin{aligned} C(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t) &\geq C_{\text{lb}}(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t)\\ C_{\text{lb}}(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t) &= \mathrm{E}\left\{\log\det\left(\boldsymbol{I} + \rho_d\boldsymbol{R}_{\boldsymbol{\nu}}^{-1}\hat{\boldsymbol{H}}_d\hat{\boldsymbol{H}}_d^H\right)\right\} \end{aligned} \tag{59}$$

where the expectation is with respect to the random variable $\hat{\boldsymbol{h}}$. The same lower bound was also proposed in [7]. As in [7], we propose a lower bound that is looser than the one given above but is simpler to handle. From (58), the matrix $\boldsymbol{R}_{\boldsymbol{\nu}}$ is a sum of three matrices. The first matrix is given by (60), shown at the bottom of the page. Each of the matrix $\mathrm{E}\{\tilde{\boldsymbol{H}}_{m_i}\tilde{\boldsymbol{H}}_{m_i}^H\}$ is a diagonal matrix, since errors in the estimates of the taps are uncorrelated. The diagonal elements are each smaller than $\operatorname{tr}\mathrm{E}\{\tilde{\boldsymbol{h}}\tilde{\boldsymbol{h}}^H\} = \frac{1}{1+\frac{c}{\sigma^2}}$. As in [7], we define a matrix $\boldsymbol{R}_{\boldsymbol{\nu}_1}$ as

$$\boldsymbol{R}_{\boldsymbol{\nu}_1} = \left(\frac{\rho_d}{1+\frac{c}{\sigma^2}} + \sigma_w^2\right)\boldsymbol{I} + \frac{1}{(L+1)\left(1+\frac{c}{\sigma^2}\right)}\boldsymbol{T}\boldsymbol{T}^H. \tag{61}$$

$$\rho_d\mathrm{E}\left\{\tilde{\boldsymbol{H}}_d\tilde{\boldsymbol{H}}_d^H\right\} = \begin{bmatrix} \rho_d\mathrm{E}\left\{\tilde{\boldsymbol{H}}_{m_1}\tilde{\boldsymbol{H}}_{m_1}^H\right\} & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \rho_d\mathrm{E}\left\{\tilde{\boldsymbol{H}}_{m_2}\tilde{\boldsymbol{H}}_{m_2}^H\right\} & & \boldsymbol{0}\\ \vdots & & \ddots & \\ \boldsymbol{0} & & & \rho_d\mathrm{E}\left\{\tilde{\boldsymbol{H}}_{m_J}\tilde{\boldsymbol{H}}_{m_J}^H\right\} \end{bmatrix}. \tag{60}$$

Since $\boldsymbol{R}_{\boldsymbol{\nu}_1} \geq \boldsymbol{R}_{\boldsymbol{\nu}}$[1] and $\boldsymbol{A} \geq \boldsymbol{B}$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ both being positive definite, implies that $|\boldsymbol{A}^{-1}| \leq |\boldsymbol{B}^{-1}|$ [6, p. 471], it follows that

$$\left|\boldsymbol{I} + \rho_d \boldsymbol{R}_{\boldsymbol{\nu}_1}^{-1} \hat{\boldsymbol{H}}_d \hat{\boldsymbol{H}}_d^H\right| \leq \left|\boldsymbol{I} + \rho_d \boldsymbol{R}_{\boldsymbol{\nu}}^{-1} \hat{\boldsymbol{H}}_d \hat{\boldsymbol{H}}_d^H\right|. \tag{62}$$

This is used to propose the lower bound

$$\begin{aligned} &C_{\text{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) \\ &\quad = \mathrm{E}\left\{\log\left|\boldsymbol{I} + \rho_d \frac{\frac{c}{\sigma^2}}{1 + \frac{c}{\sigma^2}} \boldsymbol{R}_{\boldsymbol{\nu}_1}^{-1} \bar{\boldsymbol{H}}_d \bar{\boldsymbol{H}}_d^H\right|\right\} \end{aligned} \tag{63}$$

where $\bar{\boldsymbol{H}}_d$ is obtained by normalizing $\hat{\boldsymbol{H}}_d$. Specifically, the channel $\bar{\boldsymbol{h}}$ that generates $\bar{\boldsymbol{H}}_d$ is normalized to zero mean, i.i.d. Gaussian with variance of each tap equal to $\frac{1}{L+1}$.

E. QPP Schemes

In this subsection, we introduce a family of placement schemes called QPP schemes. This family is divided into different classes based on the minimum allowable cluster size. The class of schemes for which α is the minimum cluster size is denoted as QPP-α. Intuitively, the QPP-α scheme is formed by first breaking the known symbols into as many clusters as possible each of length at least α and then placing these clusters such that the unknown symbol blocks are as "equal" as possible. We give the formal definition as follows.

Definition 1: Given α and a frame with N unknown symbols and $P \geq \alpha$ known symbols, let $J_\alpha = \lfloor \frac{P}{\alpha} \rfloor + 1$. A placement scheme $\mathcal{P} = (\boldsymbol{n}, \boldsymbol{m})$ belongs to QPP-α if and only if

1) $\boldsymbol{n} \in \mathcal{N}^{J_\alpha}$ where

$$\begin{aligned} \mathcal{N}^{J_\alpha} = \Bigg\{ &(n_1, \ldots, n_{J_\alpha+1}) \colon \sum_{i=2}^{J_\alpha} n_i = P \\ &\& \, n_1 = n_{J_\alpha+1} = L \\ &\& \, \min(\{n_2, \ldots, n_{J_\alpha}\}) \geq \alpha \Bigg\}. \end{aligned}$$

2) $\boldsymbol{m} \in \mathcal{M}^{J_\alpha}$ where

$$\begin{aligned} \mathcal{M}^{J_\alpha} = \Bigg\{ &(m_1, \ldots, m_{J_\alpha}) \colon \sum_i m_i = N \\ &\& \, m_i \in \left\{ \left\lfloor \frac{N}{J_\alpha} \right\rfloor, \left(\left\lfloor \frac{N}{J_\alpha} \right\rfloor + 1 \right) \right\} \Bigg\}. \end{aligned}$$

Any element of the set $\mathcal{N}^{J_\alpha}$ is denoted as $\boldsymbol{n}_{J_\alpha} = (\bar{n}_1, \ldots, \bar{n}_{J_\alpha})$ and similarly any element of the set $\mathcal{M}^{J_\alpha}$ is denoted as $\boldsymbol{m}_{J_\alpha} = (\bar{m}_1, \ldots, \bar{m}_{J_\alpha})$.

F. Optimality of QPP-α Schemes for Unknown Channel

We obtain optimal training $(\mathcal{P}^*, \rho_d^*, \rho_t^*, \boldsymbol{s}_t^*)$ as

$$(\mathcal{P}^*, \rho_d^*, \rho_t^*, \boldsymbol{s}_t^*) = \arg \max_{\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t} C_{\text{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t). \tag{64}$$

We first obtain an upper bound on $C_{\text{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t)$ that is a function of only ρ_d, ρ_t, N, and P.

Lemma 3:

$$\begin{aligned} &C_{\text{lb}}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t) \\ &\leq \sum_{i=1}^{J_\alpha} \mathrm{E}\Bigg\{\log\Bigg|\boldsymbol{I} + \frac{\rho_d}{\sigma_w^2} \frac{(P + L)\rho_t}{(P+L)\rho_t + (L+1)(\rho_d + \sigma_w^2)} \\ &\qquad \cdot \bar{\boldsymbol{H}}_{\bar{m}_i}^H \bar{\boldsymbol{H}}_{\bar{m}_i}\Bigg|\Bigg\} \end{aligned} \tag{65}$$

where $\boldsymbol{m}_{J_\alpha} = (\bar{m}_1, \ldots, \bar{m}_{J_\alpha})$ is the unknown symbol block length tuple for the QPP-α scheme with N unknown and P known symbols.

Proof: Refer to Appendix V.

The following theorem shows that under the assumption that $\alpha \geq 2L + 1$, the placement schemes belonging to QPP-α are optimal. Furthermore, an optimal choice of training symbols is also given.

Theorem 4: Given any energy tradeoff (ρ_d, ρ_t), under the assumption that $\alpha \geq (2L + 1)$ and $P \geq \alpha$, the placement scheme $\mathcal{P}^*$ and training $\boldsymbol{s}_t^*$ is optimal if

1) $\mathcal{P}^*$ belongs to QPP-α.

2)

$$|s_{it}^k| = \begin{cases} \sqrt{\dfrac{(P+L)\rho_t}{J-1}}, & \text{if } k = (L+1), \ i = 2, \ldots, J \\ 0, & \text{otherwise.} \end{cases} \tag{66}$$

If $(L+1) \leq P < \alpha$, the known symbols are placed at the beginning and the end of the packet such that at least $(L+1)$ are at one of the ends. That is, a placement scheme $\mathcal{P}^*$ and training symbols $\boldsymbol{s}_t^*$ are optimal if

1) $\mathcal{P} = (\boldsymbol{m}, \boldsymbol{n})$ where

$$\boldsymbol{m} = (N), \boldsymbol{n} = (2L + 1 + \beta, \, P - 1 - \beta)$$

and $0 \leq \beta \leq P - (L+1)$.

2)

$$|s_{it}^k| = \begin{cases} \sqrt{(P+L)\rho_t}, & \text{if } k = (L+1), \ i = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{67}$$

In either case we have

$$\begin{aligned} &C_{\text{lb}}(\mathcal{P}^*, \rho_d, \rho_t, \boldsymbol{s}_t^*) \\ &= \sum_{i=1}^{J_\alpha} \mathrm{E} \log\Bigg(1 + \frac{\rho_d}{\sigma_w^2} \frac{(P+L)\rho_t}{(P+L)\rho_t + (L+1)(\rho_d + \sigma_w^2)} \\ &\qquad \cdot \bar{\boldsymbol{H}}_{\bar{m}_i}^H \bar{\boldsymbol{H}}_{\bar{m}_i}\Bigg). \end{aligned} \tag{68}$$

Proof: Refer to Appendix VI.

We find that QPP-α placement schemes that were found to be optimal in the known channel scenario [12] are optimal for this scenario too. From (66) and (67), we find that for the optimal choice of training symbols, the L symbols at the beginning and the end of each known symbol cluster are zero. If these symbols are nonzero, we find that these symbols contribute additional noise to the received data because of the error in the channel estimate. Also we find that in each cluster, there is only one nonzero

[1] Given two Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we say $\boldsymbol{A} \geq \boldsymbol{B}$ if and only if the matrix $(\boldsymbol{A} - \boldsymbol{B})$ is positive semidefinite.

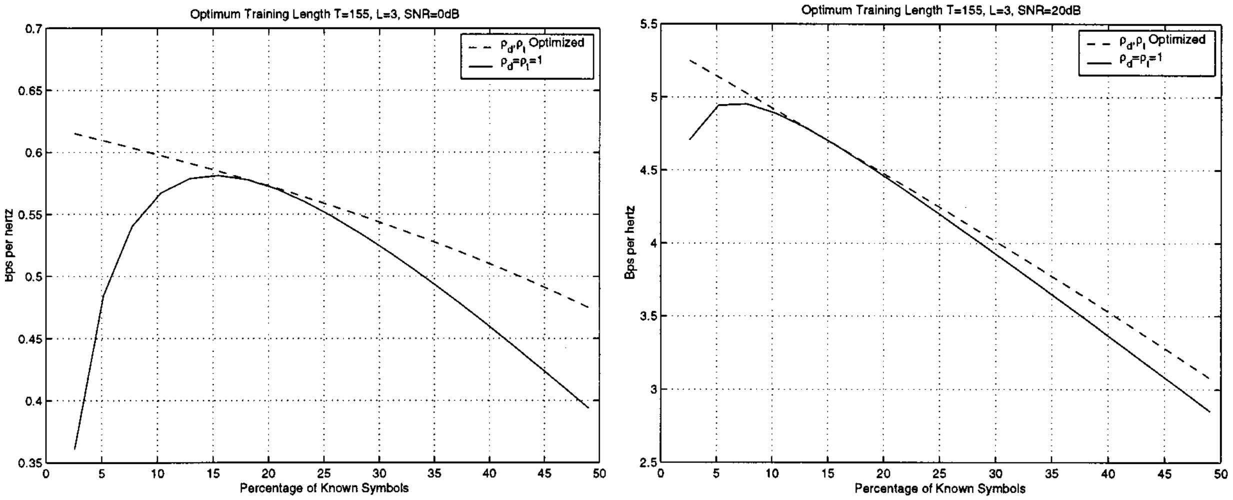

Fig. 9. Variation of lower bound with percentage of known symbols for $T = 155$ and $L = 3$ at different SNRs.

training symbol. This design makes sure that the training is always orthogonal. For $L \leq \alpha \leq 2L$, it is difficult to analytically obtain the optimal placement schemes.

The minimum known symbol cluster size α is also a design parameter. The following theorem gives the optimal value of α.

Theorem 5: For $\alpha \geq 2L+1$, $C_{\rm lb}(\mathcal{P}^*, \rho_d, \rho_t, \boldsymbol{s}_t^*)$ is a monotonically decreasing function of α.

Proof: Refer to Appendix VII.

The obtained placement schemes are optimal for any energy allocation. The following theorem gives the optimal energy allocation between training and data under the assumption that the optimal placement scheme and training symbols are used.

Theorem 6: The optimal energy distribution is given by

$$\rho_d^* = \left(\sqrt{\gamma} - \sqrt{\gamma - 1}\right) \frac{\sqrt{\gamma}}{g}$$
$$\rho_t^* = \left(\sqrt{\gamma} - \sqrt{\gamma - 1}\right) \frac{1}{\sqrt{hk}}. \tag{69}$$

where

$$h = \frac{P}{T}$$
$$g = \frac{N}{T}$$
$$k = \frac{(P+L)(N-L-1)}{(L+1)(T+N\sigma_w^2)}$$
$$\gamma = \frac{h}{k} + 1.$$

Proof: The proof is similar to the one for Theorem 3.

G. An Upper Bound on the Training-Based Capacity

We obtain an upper bound on the training-based capacity by assuming that the receiver estimates the channel perfectly from training. In other words, we assume that $\hat{\boldsymbol{h}} = \boldsymbol{h}$. Clearly, the maximum i.i.d. mutual information in this case is an upper bound on $C(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t)$. (Note that the upper bound may not be tight.) The relation between the input and output now becomes

$$\boldsymbol{y} = \boldsymbol{H}_d \boldsymbol{s}_d + \boldsymbol{w}_d. \tag{70}$$

It is easy to see that the mutual information is not a function of $\boldsymbol{s}_t$ and ρ_t. Given N and P, the upper bound $C_{\rm ub}(\mathcal{P}, \rho_d)$ is given by

$$C_{\rm ub}(\mathcal{P}, \rho_d) = \max_{f_{\rm i.i.d.}(\boldsymbol{s}_d)} I(\boldsymbol{y}_d, \boldsymbol{h}; \boldsymbol{s}_d) \tag{71}$$

where $f_{\rm i.i.d.}(\boldsymbol{s}_d)$ is an i.i.d. probability distribution satisfying the energy constraint. It is easy to see that

$$C_{\rm ub}(\mathcal{P}, \rho_d) = \mathrm{E}\left\{\log\left|\boldsymbol{I} + \frac{\rho_d}{\sigma_w^2} \bar{\boldsymbol{H}}_d^H \bar{\boldsymbol{H}}_d\right|\right\} \tag{72}$$
$$= \sum_{i=1}^{J} \mathrm{E}\left\{\log\left|\boldsymbol{I} + \frac{\rho_d}{\sigma_w^2} \bar{\boldsymbol{H}}_{m_i}^H \bar{\boldsymbol{H}}_{m_i}\right|\right\}. \tag{73}$$

We now consider optimize placement of training with respect to this upper bound. Given ρ_d, we find out the optimal placement as

$$\mathcal{P}^*(\rho_d) = \arg\max_{\mathcal{P}} C_{\rm ub}(\mathcal{P}, \rho_d). \tag{74}$$

Upon comparing (72) and (63), we note that both the lower bound and the upper bound depend on placement in exactly the same way. Hence, it can be shown that for $\alpha \geq (2L+1)$ and $P \geq \alpha$, the placement $\mathcal{P}^*(\rho_d)$ is optimal if it belongs to a QPP-α scheme. The optimal placement is therefore independent of ρ_d. We can now try to fix this placement and optimize ρ_d and ρ_t. The optimum value of ρ_t is in fact equal to zero and thus $\rho_d^* = \frac{T}{N}$.

V. Simulation

In this section, we explore the properties of training-based capacity for both OFDM and single-carrier systems through simulations. First, we present the simulations for OFDM systems followed by the simulations for the single-carrier systems. We conclude with some comparisons between the OFDM and single-carrier systems.

A. OFDM System

Fig. 9 shows the variation of lower bound given in (35) for training-based capacity with the percentage of known symbols

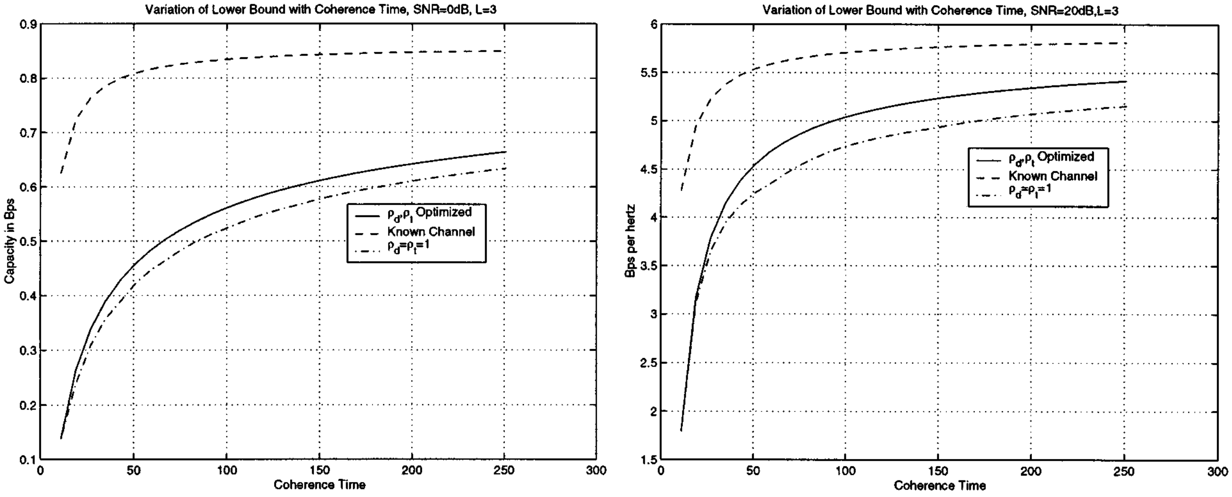

Fig. 10. Variation of lower bound with coherence interval for $L = 3$ at different SNRs with optimized P at each T.

$\frac{100P}{T}$. The coherence interval T is equal to 155. We assume that the channel is of length 4. Plots are shown for 0- and 20-dB SNR. Curves are plotted for both $\rho_d = \rho_t = 1$ and (ρ_d, ρ_t) optimized cases. We assume that the optimal placement scheme was used in all cases. We find that for the equal energy allocation case, the bound increases and then falls. The optimum percentage of known symbols is approximately equal to 15% for SNR = 0 dB and 6% for SNR = 20 dB. It is natural to expect the optimum percentage of known symbols to decrease with SNR since the quality of the estimate improves with SNR. For the optimized energy-allocation case, the bound decreases monotonically. From simulations we find that $P = (L + 1)$ is always optimal. For single-carrier systems with single known symbol cluster and optimized energy tradeoff, it is indeed true that $P = (L + 1)$ is optimal [7]. We conjecture that this can be shown to be true for OFDM systems as well. At high SNR, the gain in optimizing ρ_d, ρ_t is minimal. We also note that for the equal energy allocation scenario, the bound rises rapidly but falls at the smaller rate.

In order to evaluate the asymptotic performance of the training-based systems, we plot the variation of $C_{\rm lb}(\mathcal{P}^*, \rho_d, \rho_t, \boldsymbol{s}_t^*)$ with the coherence interval T in Fig. 10. The plots are shown for both equal energy and optimized energy allocation for both low SNR (0 dB) and high SNR (20 dB). At each value of T, we evaluate the optimum number of known symbols and calculate the lower bound by setting the number of known symbols to this value. We find that at high SNR, the capacity of training-based system approaches that of the known channel faster than at low SNR. We also note that at small values of T, the gain from optimizing ρ_d, ρ_t is minimal.

In order to judge the efficacy of training-based scheme in achieving the capacity of the unknown channel, we plot the fraction of known channel capacity achieved versus SNR (see Fig. 11). We find that at $T = 155$ and SNR = 20 dB, the capacity of the trsining-based scheme is close to that of the known channel and we can thus conclude that training-based methods achieve most of the unknown channel capacity at high SNR and large T. Similar conclusions were reached in [2], [7].

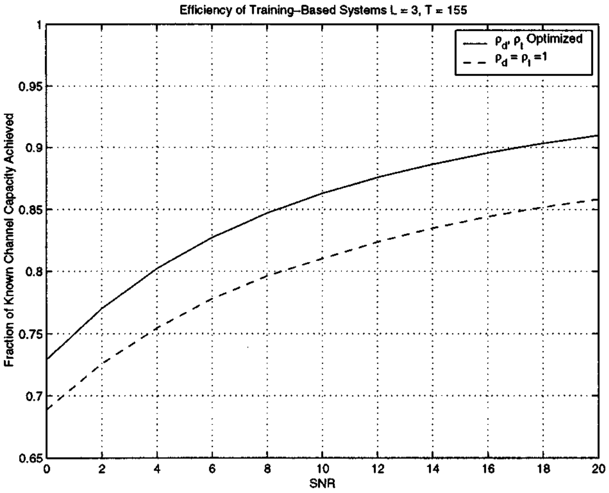

Fig. 11. Fraction of known channel capacity achieved at different SNRs for $T = 155$ and $L = 3$ with optimized P at each T.

B. Single-Carrier Systems

In this subsection, we study the training-based capacity for single-carrier systems through simulations. We evaluate the asymptotic performance of training-based systems in Fig. 12. We plot the lower bound versus the coherence interval T for low SNR (0 dB) and high SNR (20 dB). The value of L was set to 3. The minimum cluster size α was made equal to $2L + 1$. For each value of T, the optimum number of known symbols was used. The placement scheme used was a QPP-α scheme. Like in OFDM systems, we find that at high SNR, asymptotically training-based capacity approaches the known channel capacity. In order to characterize the efficiency of the training-based system with respect to SNR, we plot the fraction of known channel capacity achieved with SNR (see Fig. 13). As earlier, we find that training-based systems achieve most of the unknown channel capacity at SNR = 20 dB and $T = 155$.

C. Comparison of OFDM and Single-Carrier Systems

In this subsection, we compare the performance of OFDM systems with single-carrier systems in different scenarios.

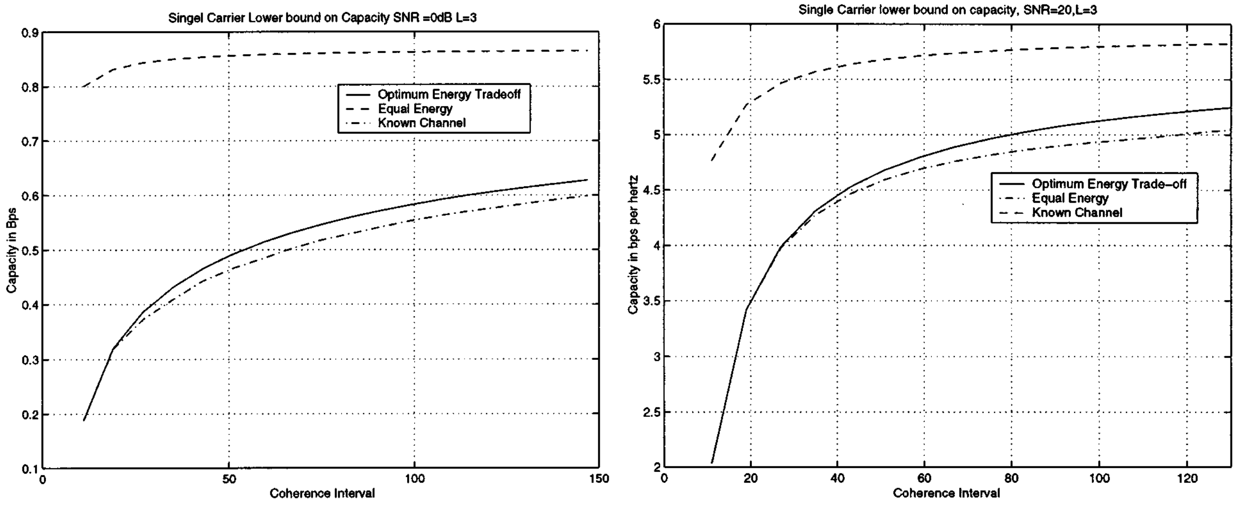

Fig. 12. Variation of lower bound with coherence interval for single-carrier systems $T = 155$ and $L = 3$ for different SNRs.

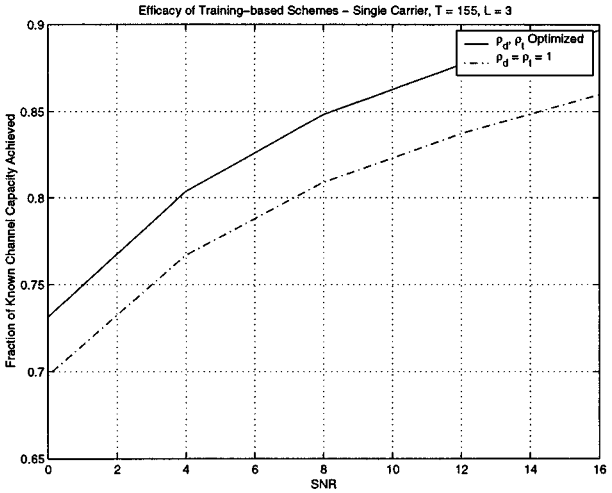

Fig. 13. Fraction of known channel capacity achieved at different SNRs for $T = 155$ and $L = 3$.

Fig. 14 compares the variation of the training-based lower bound with percentage of known symbols for OFDM and single-carrier systems with the coherence interval $T = 155$ and the channel length equal to 4. We find that the training-based capacity for single-carrier systems is consistently better than that of OFDM systems. For optimized (ρ_d, ρ_t), we find that the percentage difference is less than 5%. For equal energy case, at low SNR, we find that the single-carrier system performs considerably better than the OFDM system at small percentage of known symbols. This difference becomes smaller with the number of known symbols. At high SNR, the percentage difference between OFDM and single-carrier systems becomes much smaller.

Fig. 15 compares the variation of the training-based lower bound with the coherence time T for OFDM and single-carrier systems with the channel length equal to 4. As expected, the known channel capacity for OFDM converges to that for single-carrier systems at large T. We find that for optimized (ρ_d, ρ_t), the difference between OFDM and single-carrier systems is quite small. For equal energy allocation, though, we find that at intermediate values of T, single-carrier systems can outperform OFDM systems by as much as 10%.

VI. Conclusion

The problem of designing optimal training symbol placement schemes for block frequency-selective fading channels is presented. It is assumed that the receiver forms an MMSE estimate of the channel based on only training. The problem is addressed for both OFDM and single-carrier systems separately since the paradigm for channel estimation is different for each system. The metric used for optimization was a tight lower bound on the i.i.d. capacity of the system.

It is shown that for OFDM systems, under the assumption that the training tones are of equal energy, the optimal placement scheme is that for which the training tones are selected periodically. We also present expressions for optimal energy allocation between training and data. For single-carrier system, we assume that the known symbols are placed in clusters of length $\alpha \geq 2L + 1$. For $\alpha \geq (2L + 1)$, we show that the placement schemes belonging to the QPP-α family are optimal. Furthermore, a choice of optimal training symbols is presented. Expressions for optimal energy allocation between data and training are given.

From simulations, we find that at large values of T and at high SNR, training-based systems achieve most of the unknown channel capacity. At low SNR, however, this is not true. The comparison of the lower bound for OFDM and single-carrier systems shows that the single-carrier system performs better than the OFDM systems. This is to be expected because the OFDM system drops some received data for simpler receiver implementation. We find that for optimal energy allocation, the percentage difference between the two systems is quite small. For equal energy case, on the other hand, the single-carrier system might be considerably better than the OFDM system for some values of T and P.

We list some related issues that are beyond the scope of this paper but have both theoretical and practical interest. In this paper, we assume that the channel taps are i.i.d. A more realistic assumption is to let channel taps be correlated and not

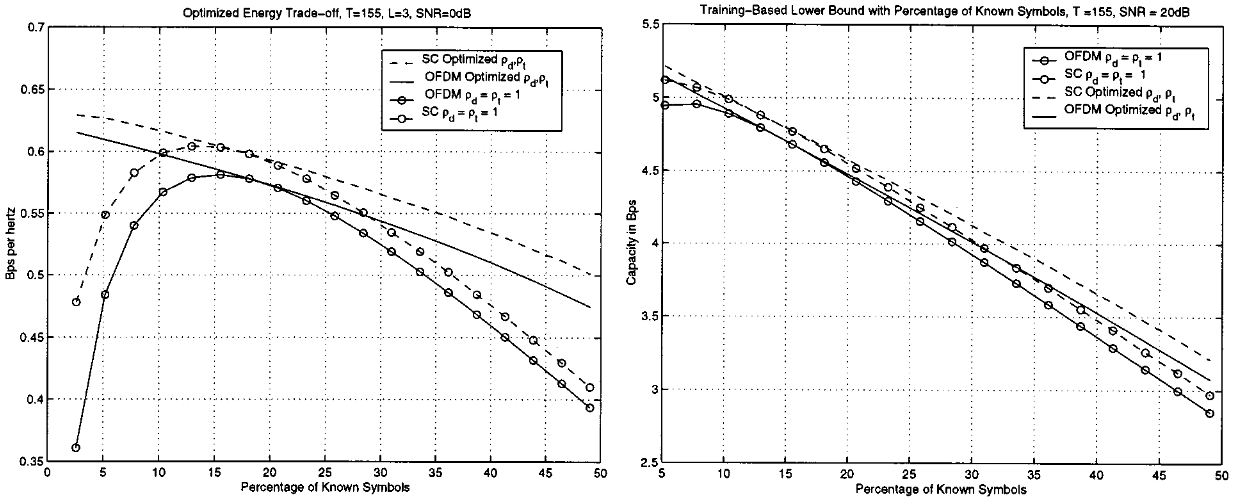

Fig. 14. Comparison of the variation of training-based lower bound with the percentage of known symbols for OFDM and single-carrier systems with at $T = 155$ and $L = 3$ for different SNR's.

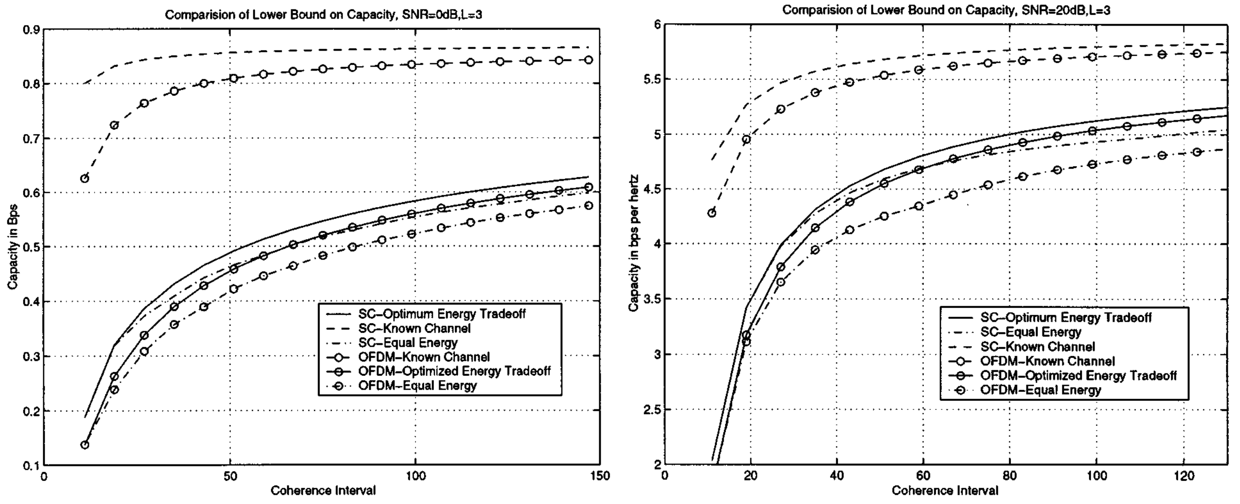

Fig. 15. Comparison of the variation of training-based lower bound with the coherence interval T for OFDM and single-carrier systems for $L = 3$ at different SNRs.

necessarily identically distributed. This model turns out to be quite difficult to analyze. Nevertheless, it is definitely a interesting problem. The extension of the single-carrier results for equal energy training is also an open problem. The extensions of these placement schemes to multiple antenna systems is an interesting research topic. Another interesting problem is optimizing training for receivers that assume that the channel estimate is perfect.

Appendix I
Proof of Theorem 1

The following proof is similar to the one in [2]. Note that $\boldsymbol{n}_d \sim \mathcal{CN}(0, \boldsymbol{R}_{\boldsymbol{\nu}_d})$ belongs to $\Omega(f_{\text{i.i.d.}}(\boldsymbol{s}_d))$ for every $f_{\text{i.i.d.}}(\boldsymbol{s}_d)$. It can be seen that

$$C_{\text{lb}} \le \sup_{f_{\text{i.i.d.}}(\boldsymbol{s}_d)} I\left(\boldsymbol{D}_d\boldsymbol{s}_d + \boldsymbol{n}_d,\ \hat{\boldsymbol{D}}_d;\ \boldsymbol{s}_d\right) \tag{75}$$

where $\boldsymbol{n}_d \sim \mathcal{CN}(0, \boldsymbol{R}_{\boldsymbol{\nu}_d})$. Therefore, we have

$$C_{\text{lb}} \le \text{E}\log\det\left(\boldsymbol{I} + \rho_d\boldsymbol{R}_{\boldsymbol{\nu}}^{-1}\hat{\boldsymbol{D}}_d\hat{\boldsymbol{D}}_d^H\right). \tag{76}$$

We next obtain a lower bound on C_{lb} by fixing $\boldsymbol{s}_d \sim \mathcal{CN}(0, \rho_d\boldsymbol{I})$ and then taking the infimum among the distributions in $\Omega(f_{\text{i.i.d.}}(\boldsymbol{s}_d))$. We then know that the worst case distribution is independent Gaussian [2]. Therefore,

$$C_{\text{lb}} \ge \text{E}\log\det\left(\boldsymbol{I} + \rho_d\boldsymbol{R}_{\boldsymbol{\nu}}^{-1}\hat{\boldsymbol{D}}_d\hat{\boldsymbol{D}}_d^H\right) \tag{77}$$

where the expectation is with respect to $\hat{\boldsymbol{D}}$. From (77) and (76) we have the theorem.

Appendix II
Proof of Lemma 1

We have

$$C_{\text{lb}}(\mathcal{P},\ \rho_d,\ \rho_t) = \sum_{i=1}^{N} f\left(\frac{1-k_i}{\gamma_d+k_i}\right) \tag{78}$$

$$\le Nf\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1-k_i}{\gamma_d+k_i}\right) \tag{79}$$

$$= Nf\left(-1+\frac{1+\gamma_d}{N}\sum_{i=1}^{N}\frac{1}{\gamma_d+k_i}\right) \tag{80}$$

$$\le Nf\left(-1+\frac{1+\gamma_d}{N}\max_{\mathcal{P}}\sum_{i=1}^{N}\frac{1}{\gamma_d+k_i}\right). \tag{81}$$

The first inequality holds because the function $f(\cdot)$ is concave. The second inequality follows because $f(\cdot)$ is monotonically decreasing. □

Appendix III
Proof of Lemma 2

We define the metric $M(\mathcal{P}, \rho_d, \rho_t) = \sum_{i=1}^{N}\frac{1}{c+k_i}$. From Lemma 1, we have that

$$C_{\rm lb}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t^*) \le Nf\left(-1+\frac{1+\gamma_d}{N}\max_{\mathcal{P},\rho_d,\rho_t} M(\mathcal{P}, \rho_d, \rho_t)\right). \tag{82}$$

From (25), we have

$$M(\mathcal{P}, \rho_d, \rho_t) = \sum_{i=1}^{N}\frac{1}{\frac{\boldsymbol{p}_i\boldsymbol{V}_L}{\sqrt{L+1}}\left(\gamma_d\boldsymbol{I}+\left(\boldsymbol{I}+\frac{1}{\gamma_t}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\right)\frac{\boldsymbol{V}_L^H\boldsymbol{p}_i^H}{\sqrt{L+1}}} \tag{83}$$

where $\boldsymbol{V}_L = \sqrt{N+P}\boldsymbol{W}_L$, $\boldsymbol{V}_{\mathcal{P}L} = \sqrt{N+P}\,\boldsymbol{W}_{\mathcal{P}L}$, and γ_t is the inverse training SNR $\frac{\sigma^2}{\rho_t}$. By the Cauchy–Schwartz inequality,[2] we have

$$M(\mathcal{P}, \rho_d, \rho_t) \le \sum_{i=1}^{N}\frac{\boldsymbol{p}_i\boldsymbol{V}_L}{\sqrt{L+1}}\left(\gamma_d\boldsymbol{I}+\left(\boldsymbol{I}+\frac{1}{\gamma_t}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\right)^{-1}\frac{\boldsymbol{V}_L^H\boldsymbol{p}_i^H}{\sqrt{L+1}} \tag{84}$$

$$= \frac{1}{\gamma_d(L+1)}\sum_{i=1}^{N}\boldsymbol{p}_i\boldsymbol{V}_L \cdot\left(\boldsymbol{I}-\left((\gamma_d+1)\boldsymbol{I}+\frac{\gamma_d}{\gamma_t}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\right)\boldsymbol{V}_L^H\boldsymbol{p}_i^H \tag{85}$$

$$= \frac{N}{\gamma_d}-\frac{1}{\gamma_d(L+1)}\sum_{i=1}^{N}\boldsymbol{p}_i\boldsymbol{V}_L \cdot\left((\gamma_d+1)\boldsymbol{I}+\frac{\gamma_d}{\gamma_t}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\boldsymbol{V}_L^H\boldsymbol{p}_i^H \tag{86}$$

$$= \frac{N}{\gamma_d}-\frac{1}{\gamma_d(L+1)} \cdot{\rm tr}\left(\boldsymbol{V}_L\left((\gamma_d+1)\boldsymbol{I}+\frac{\gamma_d}{\gamma_t}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\boldsymbol{V}_L^H\right) + \frac{1}{\gamma_d(L+1)}\sum_{i=1}^{P}\boldsymbol{q}_i\boldsymbol{V}_L \cdot\left((\gamma_d+1)\boldsymbol{I}+\frac{\gamma_d}{\gamma_t}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\boldsymbol{V}_L^H\boldsymbol{q}_i^H \tag{87}$$

where $\boldsymbol{q}_i$ is a unit row vector with a 1 in the index of the ith training tone and 0's elsewhere. Equation (85) follows from the application of the matrix inversion lemma.[3] Now using some simple manipulations, the above can be rewritten as

$$= \frac{N}{\gamma_d}-\frac{(N+P)}{\gamma_d(\gamma_d+1)(L+1)}{\rm tr}\left(\boldsymbol{I}+\frac{\gamma_d}{\gamma_t(\gamma_d+1)}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1} + \frac{1}{\gamma_d(\gamma_d+1)(L+1)} \cdot{\rm tr}\left(\boldsymbol{V}_{\mathcal{P}L}\left(\boldsymbol{I}+\frac{\gamma_d}{\gamma_t(\gamma_d+1)}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1}\boldsymbol{V}_{\mathcal{P}L}^H\right) \tag{88}$$

$$= \frac{N}{\gamma_d}-\frac{(N+P)}{\gamma_d(\gamma_d+1)(L+1)}{\rm tr}\left(\boldsymbol{I}+\frac{\gamma_d}{\gamma_t(\gamma_d+1)}\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right)^{-1} + \frac{P\gamma_t}{\gamma_d^2(L+1)}-\frac{\gamma_t}{\gamma_d^2(L+1)}{\rm tr}\left(\boldsymbol{I}+\frac{\gamma_d}{\gamma_t(\gamma_d+1)}\boldsymbol{V}_{\mathcal{P}L}\boldsymbol{V}_{\mathcal{P}L}^H\right)^{-1} \tag{89}$$

$$= \frac{N}{\gamma_d}+\frac{\gamma_t}{\gamma_d^2}-\left(\frac{(N+P)}{\gamma_d(\gamma_d+1)(L+1)}+\frac{\gamma_t}{\gamma_d^2(L+1)}\right)\sum_{i=1}^{L+1} \cdot\frac{1}{\lambda_i+1} \tag{90}$$

where $\{\lambda_i\}_{i=1}^{L+1}$ are the eigenvalues of the matrix $\frac{\gamma_d}{\gamma_t(\gamma_d+1)} \cdot \boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}$. Equation (89) follows from the matrix inversion lemma. Equation (90) follows from the fact that $\boldsymbol{V}_{\mathcal{P}L}\boldsymbol{V}_{\mathcal{P}L}^H$ has only $(L+1)$ nonzero eigenvalues and they are the same as those of $\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}$. We now note that

$$\sum_{i=1}^{L+1}\lambda_i = \frac{\gamma_d}{\gamma_t(\gamma_d+1)}{\rm tr}\left(\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}\right) = \frac{\gamma_d P(L+1)}{\gamma_t(\gamma_d+1)}. \tag{91}$$

Under the constraint (91), it is easy to see that

$$\sum_{i=1}^{L+1}\frac{1}{\lambda_i+1} \ge \frac{(L+1)}{1+\frac{P\gamma_d}{\gamma_t(\gamma_d+1)}} \tag{92}$$

with equality if and only if all the λ_i are equal or, equivalently, the matrix $\boldsymbol{V}_{\mathcal{P}L}^H\boldsymbol{V}_{\mathcal{P}L}$ must be equal to a constant times identity. Combining (90) and (92), we have

$$M(\mathcal{P}, \rho_d, \rho_t) \ge \frac{N}{\gamma_d+\frac{1}{1+\frac{P}{\gamma_t}}}. \tag{93}$$

We then have that

$$C_{\rm lb}(\mathcal{P}, \rho_d, \rho_t, \boldsymbol{s}_t^*) \le Nf\left(-1+\frac{1+\gamma_d}{\gamma_d+\frac{1}{1+\frac{P}{\gamma_t}}}\right) = Nf\left(\frac{P}{\gamma_t(\gamma_d+1)+P\gamma_d}\right). \tag{94}$$

□

Appendix IV
Proof of Theorem 3

The objective is to maximize

$$\rho_{eff} \triangleq \frac{P\rho_t\rho_d}{\sigma_w^2(P\rho_t+(L+1)(\rho_d+\sigma_w^2))} \tag{95}$$

[2] If $\boldsymbol{x}$ is a unit norm row vector and $\boldsymbol{A}$ is a matrix then $\frac{1}{\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^H} \le \boldsymbol{x}\boldsymbol{A}^{-1}\boldsymbol{x}^H$. See, e.g., [6].

[3] $(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\,(\boldsymbol{C}^{-1}+\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}$. See, e.g., [5].

under the power constraint

$$\frac{N\rho_d + P\rho_t}{N+P} = 1.$$

This is a simple optimization problem similar to one performed in [2], [7]. □

APPENDIX V
PROOF OF LEMMA 3

We have

$$C_{\mathrm{lb}}(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t) = \mathrm{E}\left\{\log\left|\boldsymbol{I}+\rho_d \, \frac{\frac{c}{\sigma^2}}{1+\frac{c}{\sigma^2}} \, \boldsymbol{R}_{\boldsymbol{\nu}_1}^{-1}\bar{\boldsymbol{H}}_d\bar{\boldsymbol{H}}_d^H\right|\right\} \tag{96}$$

where

$$\boldsymbol{R}_{\boldsymbol{\nu}_1} = \left(\frac{\rho_d}{1+\frac{c}{\sigma^2}} + \sigma_w^2\right)\boldsymbol{I} + \boldsymbol{T}\boldsymbol{T}^H. \tag{97}$$

We obtain an upper bound on $C_{\mathrm{lb}}(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t)$ that is a function of only the energy tradeoff $(\rho_d,\, \rho_t)$. We have

$$\begin{aligned}
&C_{\mathrm{lb}}(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t)\\
&\quad\le \mathrm{E}\left\{\log\left|\boldsymbol{I} + \frac{\rho_d}{\sigma_w^2}\, \frac{c}{c+(L+1)(\rho_d+\sigma_w^2)}\, \bar{\boldsymbol{H}}_d\bar{\boldsymbol{H}}_d^H\right|\right\}\\
&\quad\le \mathrm{E}\left\{\log\left|\boldsymbol{I}+\frac{\rho_d}{\sigma_w^2}\, \frac{(P+L)\rho_t}{(P+L)\rho_t+(L+1)(\rho_d+\sigma_w^2)}\, \bar{\boldsymbol{H}}_d^H\bar{\boldsymbol{H}}_d\right|\right\}\\
&\quad= \sum_{i=1}^{J} \mathrm{E}\left\{\log\left|\boldsymbol{I}+\frac{\rho_d}{\sigma_w^2}\, \frac{(P+L)\rho_t}{(P+L)\rho_t+(L+1)(\rho_d+\sigma_w^2)}\right.\right.\\
&\qquad\qquad \left.\left.\cdot\, \bar{\boldsymbol{H}}_{m_i}^H\bar{\boldsymbol{H}}_{m_i}\right|\right\}\\
&\quad\triangleq \sum_{i=1}^{J} g(\rho_d,\, \rho_t,\, m_i).
\end{aligned}$$

The first inequality follows because[4]

$$\boldsymbol{R}_{\boldsymbol{\nu}_1} \ge \left(\frac{\rho_d}{1+\frac{c}{\sigma^2}} + \sigma_w^2\right)\boldsymbol{I}$$

and $\boldsymbol{A} \ge \boldsymbol{B}$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ both being positive definite, implies that $|\boldsymbol{B}^{-1}| \ge |\boldsymbol{A}^{-1}|$ [6, p. 471]. The second inequality follows from (49) and the fact that $|\boldsymbol{I}+\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{I}+\boldsymbol{B}\boldsymbol{A}|$. The matrices $\bar{\boldsymbol{H}}_{m_i}^H\bar{\boldsymbol{H}}_{m_i}$ are positive definite and Toeplitz. This can be used to show that the function $g(\cdot)$ has the property [12]

$$\begin{aligned}
2g(\rho_d,\, \rho_t,\, n) \ge g(\rho_d,\, \rho_t,\, n+k) + g(\rho_d,\, \rho_t,\, n-k),&\\
\forall n \in \mathcal{Z}^+ \,\&\, k \in \{0,\, 1,\, \ldots,\, n\}.&
\end{aligned} \tag{98}$$

It is easy to see that the above property implies that given $J = J_\alpha$

$$\sum_{i=1}^{J} g(\rho_d,\, \rho_t,\, m_i) \le \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, \bar{m}_i) \tag{99}$$

where $\boldsymbol{m}_{J_\alpha} = (\bar{m}_1,\, \ldots,\, \bar{m}_{J_\alpha})$ is the unknown symbol block length tuple for the QPP-α scheme with N unknown and P known symbols [12]. If $J < J_\alpha$ then

$$\begin{aligned}
\sum_{i=1}^{J} g(\rho_d,\, \rho_t,\, m_i) &= \sum_{i=1}^{J} g(\rho_d,\, \rho_t,\, m_i) + \underbrace{\sum_{i=J+1}^{J_\alpha} g(\rho_d,\, \rho_t,\, 0)}_{0}\\
&= \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, m_i')\\
&\le \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, \bar{m}_i).
\end{aligned} \tag{100}$$

where $m_i' = m_i$ for $i = 1,\, \ldots,\, J$ and $m_i' = 0$ for $i = (J+1),\, \ldots,\, J_\alpha$. The inequality follows from (99). The properties (99) and (100) together with $J \le J_\alpha$ imply that

$$\sum_{i=1}^{J} g(\rho_d,\, \rho_t,\, m_i) \le \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, \bar{m}_i). \tag{101}$$

Finally, we note that under the constraint that each known symbol cluster is at least as big as α, the number of unknown data blocks $J \le J_\alpha$. Hence

$$C_{\mathrm{lb}}(\mathcal{P},\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t) \le \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, \bar{m}_i). \qquad \square$$

APPENDIX VI
PROOF OF THEOREM 4

We first assume that $\alpha \ge (2L+1)$ and $P \ge \alpha$. Let the placement scheme $\mathcal{P}^*$ belong to QPP-α. The particular choice of training symbols $\boldsymbol{s}_t^*$ implies that every packet starts with exactly L zeros. Moreover, each known symbol cluster starts and ends with L zeros. Each known symbol cluster has only one nonzero training symbol. The energy in training is divided equally among all these symbols. For this choice of training, it is easy to see that the matrix $\boldsymbol{T}$ as defined in (52) is equal to zero. Further $\boldsymbol{S}_t^H\boldsymbol{S}_t = (P+L)\rho_t\boldsymbol{I}$. This implies that matrix

$$\boldsymbol{R}_{\boldsymbol{\nu}_1} = \frac{\rho_d}{1+\frac{(P+L)\rho_t}{\sigma^2}}\boldsymbol{I}$$

and the lower bound can be easily evaluated as

$$C_{\mathrm{lb}}(\mathcal{P}^*,\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t^*) = \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, \bar{m}_i). \tag{102}$$

From Lemma 3, we can conclude that the choice of the placement scheme and training symbols is optimal. The proof for the case when $(L+1) \le P < \alpha$ is similar.

APPENDIX VII
PROOF OF THEOREM 5

For $\alpha \ge 2L+1$, we have

$$C_{\mathrm{lb}}(\mathcal{P}^*,\, \rho_d,\, \rho_t,\, \boldsymbol{s}_t^*) = \sum_{i=1}^{J_\alpha} g(\rho_d,\, \rho_t,\, \bar{m}_i). \tag{103}$$

[4] Given two Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we say $\boldsymbol{A} \ge \boldsymbol{B}$ if and only if the matrix $(\boldsymbol{A} - \boldsymbol{B})$ is positive semidefinite.

Hence, the lower bound depends on α only through the value of J_α. It is easy to see that J_α increases as α decreases. Given $\alpha_1 > \alpha_2$ such that $J_{\alpha_2} > J_{\alpha_1}$, we have

$$\sum_{i=1}^{J_{\alpha_1}} g(\rho_d,\, \rho_t,\, \bar{m}_i) < \sum_{i=1}^{J_{\alpha_2}} g(\rho_d,\, \rho_t,\, \bar{m}_i). \tag{104}$$

This follows from (101). □

ACKNOWLEDGMENT

The authors wish to thank the Associate Editor and the anonymous reviewers for their detailed comments which have improved the presentation of this paper. They would also like to thank the Associate Editor particularly for pointing out that there is no loss in considering linear MMSE estimators if the channel is assumed to be Gaussian.

REFERENCES

[1] A. Lapidoth and S. Shamai (Shitz), “Fading channels: How perfect need perfect side information be?,” in *Proc. IEEE Information Theory Workshop*, Kruger National Park, South Africa, June 1999, pp. 36–38.

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

[3] E. Biglieri, J. Proakis, and S. Shamai (Shitz), “Fading channels: Information-theoretic and communications aspects,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2619–2692, Oct. 1998.

[4] I. C. Abou Faycal, M. D. Trott, and S. Shamai (Shitz), “The capacity of discrete-time Rayleigh fading channels,” in *Proc. Int. Symp. Information Theory*, Ulm, Germany, June 1997, p. 473.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1990.

[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1985.

[7] H. Vikalo, B. Hassibi, B. Hochwald, and T. Kailath, “Optimal training for frequency-selective fading channels,” in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001, pp. 2105–2108.

[8] G. D. Forney, Jr. and G. Ungerboeck, “Modulation and coding for linear Gaussian channels,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2596–2618, Oct. 1998.

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

[10] R. Negi and J. Cioffi, “Pilot tone selection for channel estimation in a mobile OFDM system,” *IEEE Tran. Consumer Electron.*, vol. 44, pp. 1122–1128, Aug. 1998.

[11] S. Adireddy and L. Tong, “Detection with embedded known symbols: Optimal symbol placement and equalization,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP’00)*, vol. 5, Istanbul, Turkey, June 2000, pp. 2541–2544.

[12] ——, “Optimal placement of known symbols for nonergodic broadcast channels,” *IEEE Trans. Inform. Theory*. Also [Online]. Available: http://www.acsp.ece.cornell.edu, submitted for publication.

[13] S. Shamai (Shitz) and R. Laroia, “The intersymbol interference channel: Lower bounds on capacity and channel precoding loss,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1388–1404, Sept. 1996.

[14] S. Ohno and G. B. Giannakis, “Optimal training and redundant precoding for block transmissions with application to wireless OFDM,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, May 2001, pp. 2389–2392.

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