Optimal Training for Block Transmissions Over Doubly Selective Wireless Fading Channels

Xiaoli Ma, Georgios B. Giannakis, *Fellow, IEEE*, and Shuichi Ohno, *Member, IEEE*

***Abstract*—High data rates give rise to frequency-selective propagation, whereas carrier frequency-offsets and mobility-induced Doppler shifts introduce time-selectivity in wireless links. To mitigate the resulting time- and frequency-selective (or doubly selective) channels, optimal training sequences have been designed only for special cases: pilot symbol assisted modulation (PSAM) for time-selective channels and pilot tone-assisted orthogonal frequency division multiplexing (OFDM) for frequency-selective channels. Relying on a basis expansion channel model, in this paper, we design low-complexity optimal PSAM for block transmissions over doubly selective channels. The optimality in designing our PSAM parameters consists of maximizing a tight lower bound on the average channel capacity that is shown to be equivalent to the minimization of the minimum mean-square channel estimation error. Numerical results corroborate our theoretical designs.**

***Index Terms*—Doubly selective channels, frequency selective, mutual information, optimal training, pilot symbol assisted modulation, time-selective, wireless fading channels.**

I. Introduction

HIGH data rate wireless and mobile links suffer from time- and frequency-selective propagation effects. Mitigating these effects enables efficient transmission over such doubly selective channels and has justifiably received increasing attention over the last decade [11]. These fading channels are challenging to mitigate, but once acquired, they offer joint multipath-Doppler diversity gains [18], [24]. The quality of channel acquisition has a major impact on the overall system performance, especially when the channels are fast fading. Reliable estimation of doubly selective channels is thus well motivated.

Two classes of methods are available for the receiver to acquire channel state information (CSI): One is based on training symbols that are *a priori* known to the receiver, whereas the other relies only on the received symbols to acquire CSI blindly. Relative to training, blind schemes typically require longer data records and entail higher complexity [8], [25], [33]. Adaptive or decision-directed methods offer reduced complexity alternatives, but they are prone to error propagation, and their application is limited to slowly varying channels [7], [26]. Albeit suboptimal and bandwidth consuming, training methods remain attractive in practice because they decouple symbol detection from channel estimation, which reduces complexity and relaxes the required identifiability conditions [20].

For time-invariant channels, a training sequence is usually sent at the beginning of each transmission burst, but when the channel is time-selective, this preamble-based training method may not work well. This motivates periodic insertion of training symbols during the transmission, which is known as pilot symbol aided modulation (PSAM) [5]. PSAM is not only useful for time-selective channels but also for frequency-selective and even doubly selective channels [11], [21], [28] as well. The number and placement of pilots affects not only the quality of CSI acquisition but the transmission rate as well. Within the general class of doubly selective channels, PSAM has been optimized based on several criteria, but only for special channel models.

Optimization of PSAM for frequency-selective channels has relied on either average channel capacity bounds [1], [20], [21], [30], or the Cramér–Rao bound (CRB) of the adopted channel estimator [9], [21]. PSAM for time-selective fading channels has been designed by minimizing the channel mean-square estimation error [5] and recently by optimizing an average capacity bound [22]. PSAM for time- and frequency-selective channels has been also considered (but not optimized) in [11], [14], [28], and [32]. Specifically, the PSAM developed in [11] and [32] applies to a limited class of quasistatic channels (obeying the "snapshot" assumption [11]), whereas the statistical channel estimator in [28] requires long data records, has rather high complexity, and may suffer from error propagation effects. This paper's objective is to optimally design PSAM for doubly selective channels by capitalizing on a parsimoniously parameterized basis expansion channel model that was originally introduced in [12], [26], and [27] and more recently utilized by [3], [4], [18], [21], and [24].

The rest of the paper is organized as follows. Section II introduces the system model. Section III focuses on channel estimation and its decoupling from symbol detection. The relationship between channel estimation error and the lower bound on average capacity is the subject of Section IV. Section V deals with the design of the optimal training strategy that maximizes the lower bound on average capacity. Section VI provides a time-frequency sampling interpretation of the optimal design and specializes it to two important cases: time-selective and frequency-selective channels. Numerical examples are presented in Section VII, and Section VIII concludes the paper.

Notation: Upper (lower) bold-face letters will be used for matrices (column vectors). Superscript $\mathcal{H}$ will denote Hermi-

Manuscript received November 11, 2001; revised November 19, 2002. The work in this paper was supported by the National Science Foundation under Grant 0122431 and the Army Research Laboratory/CTA under Grant DAAD19-01-2-011. The associate editor coordinating the review of this paper and approving it for publication was Prof. Xiaodong Wang.

X. Ma and G. B. Giannakis are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: xiaoli@ece.umn.edu; georgios@ece.umn.edu).

S. Ohno is with the Department of Artificial Complex Systems Engineering, Hiroshima University Hiroshima, Japan (e-mail: o.shuichi@ieee.org).

Digital Object Identifier 10.1109/TSP.2003.810304

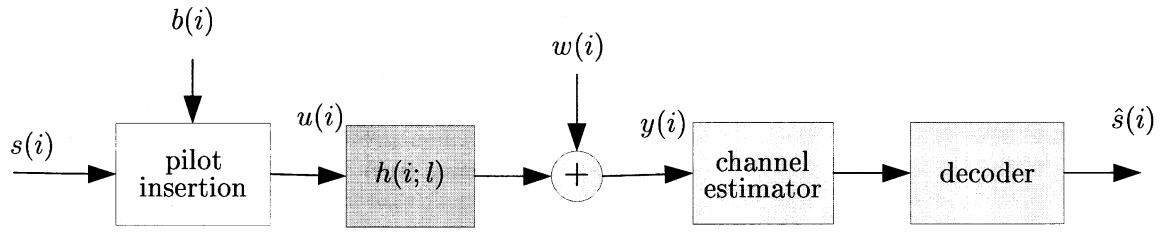


Fig. 1. Discrete-time baseband equivalent system model.

tian, $*$ conjugate, T transpose, and $\dagger$ matrix pseudoinverse. We will reserve $\star$ for convolution, $\otimes$ for Kronecker product, $\lceil\cdot\rceil$ for integer ceiling, $\lfloor\cdot\rfloor$ for integer floor, and $E[\cdot]$ for expectation with respect to all the random variables within the brackets. We will use $[\mathbf{A}]_{k,m}$ to denote the (k, m)th entry of a matrix $\mathbf{A}$, tr$(\mathbf{A})$ for its trace, and $[\mathbf{x}]_m$ to denote the mth entry of the column vector $\mathbf{x}$; finally, diag$[\mathbf{x}]$ will stand for a diagonal matrix with $\mathbf{x}$ on its main diagonal.

II. System Model

In this section, we will first present our time- and frequency-selective channel model, and then we will introduce the transmission design.

A. Time- and Frequency-Selective Channel Model

Let $h(t;\ \tau)$ denote the time-varying impulse response of our channel that includes transmit-receive filters as well as doubly selective propagation effects. With $H(f;\ \tau)$ denoting the Fourier transform of $h(t;\ \tau)$, let us also define the delay-spread $\tau_{\max}$ and the Doppler-spread $f_{\max}$ as the thresholds for which $|H(f;\ \tau)| \approx 0$ for $|\tau| > \tau_{\max}$ or $|f| > f_{\max}$. We will take the sampling period at the receiver equal to the symbol period T_s, and we will consider time intervals of NT_s s, corresponding to blocks containing N symbols each. Over each interval, say the kth, we will represent $h(t;\ \tau)$ for $t \in [kNT_s, (k+1)NT_s)$ using a) $Q+1$ coefficients $\{h_q\}_{q=0}^{Q}$ that remain invariant per block but are allowed to change with k and b) $Q+1$ Fourier bases that capture the time variation but are common $\forall\, k$. Using the serial index i, we can describe the block index as $k := \lfloor i/N \rfloor$ and write our *discrete-time baseband equivalent* channel model as (see [18] for detailed derivations):

$$h(i;\ l) = \sum_{q=0}^{Q} h_q(\lfloor i/N \rfloor;\ l) e^{j\omega_q i}, \qquad l \in [0,\ L] \tag{1}$$

where $\omega_q := 2\pi(q - Q/2)/N$, $L := \lfloor \tau_{\max}/T_s \rfloor$, and $Q := 2\lceil f_{\max} N T_s \rceil$. Because both $\tau_{\max}$ and $f_{\max}$ can be measured experimentally in practice, we assume the following.

A1) *Parameters* $\tau_{\max}$, $f_{\max}$ (and thus L, Q) are bounded, known, and satisfy $2f_{\max}\tau_{\max} < 1$.

The product $2f_{\max}\tau_{\max}$ is called delay-Doppler spread factor and plays an important role in estimating doubly selective channels. Underspread systems satisfy $2f_{\max}\tau_{\max} < 1$, which, intuitively speaking ,bounds the channel's degrees of freedom and renders channel estimation well-posed [15], [16], [18]. In fact, most ionospheric- and tropospheric-scattering, as well as other radio channels, all give rise to underspread channels; see, e.g., [23, p. 816].

Per block of N symbols, the basis expansion model (BEM) in (1) can be viewed either as deterministic or as the realization of a stochastic process with random coefficients $h_q(\lfloor i/N \rfloor;\ l)$. When transmissions experience rich scattering, and no line-of-sight is present, one can appeal to the central limit theorem to validate the following assumption in the random viewpoint:

A2) The BEM coefficients $h_q(\lfloor i/N \rfloor;\ l)$ are zero-mean, complex Gaussian random variables with variance $\sigma^2_{q,l}$.

The BEM offers a parsimonious finite-parameter representation of doubly selective channels and was originally introduced in [12], [26], and [27]. The BEM in [26, eq. (2)] or [27, eq. (4)] was expressed as

$$h(i;\ l) = \sum_{q=0}^{Q} h_q(\lfloor i/N \rfloor;\ l) f_q(i), \qquad l \in [0,\ L].$$

Specific choices for $\{f_q(i)\}_{q=0}^{Q}$ have included polynomial, wavelet, or, Fourier bases [3], [4], [12], [19]. The canonical model in [3] and [24] corresponds to substituting $f_q(i) := e^{j2\pi iq/N}$. The BEM considered here is also FFT-based but differs from the canonical model in the following major aspects: i) [3] and [24] focus on spread-spectrum transmissions and assume that the channel varies from symbol to symbol, which we do not have to do, and ii) [3] and [24] consider serial transmissions through a continuous-time scalar channel; we work with a discrete-time baseband equivalent matrix-vector channel model that is suitable for block transmissions.

B. Block Transmission Design

Fig. 1 depicts a general discrete-time baseband equivalent transmission format when communicating through the doubly selective channel (1). Two types of sub-blocks can be identified in each transmitted block: One type contains the information symbols, whereas the other includes the training (or pilot) symbols. We use two arguments (n and k) to describe the serial index $i = kN + n$ for $n \in [0,\ N-1]$, and denote the $(n+1)$st entry of the kth block as $[\mathbf{u}(k)]_n := u(kN+n)$. Each block $\mathbf{u}(k)$ includes N_s information symbols $\mathbf{s}(k) := [s(kN_s), \ldots, s(kN_s + N_s - 1)]^T$, and N_b training symbols $\mathbf{b}(k) = [b(kN_b), \ldots, b(kN_b + N_b - 1)]^T$, which are known to both transmitter and receiver.

After parallel to serial (P/S) multiplexing, the blocks $\mathbf{u}(k)$ are transmitted through the time- and frequency-selective channel $h(i;\ l)$ modeled as in (1). The ith received sample can be written as

$$y(i) = \sum_{l=0}^{L} h(i;\ l) u(i-l) + w(i) \tag{2}$$

where $w(i)$ is additive white Gaussian noise (AWGN) with mean zero and variance σ_w^2.

We will find it convenient to work with a block-form of the BEM we construct, after serial to parallel (S/P) conversion, by collecting the samples $y(i)$ into $N \times 1$ blocks: $\mathbf{y}(k) = [y(kN), y(kN+1), \ldots, y(kN+N-1)]^T$. Selecting also $N \geq L$, we can write the matrix-vector counterpart of (2) as

$$\mathbf{y}(k) = \mathbf{H}(k)\mathbf{u}(k) + \mathbf{H}^{\text{ibi}}(k)\mathbf{u}(k-1) + \mathbf{w}(k) \qquad (3)$$

where $\mathbf{w}(k) := [w(kN), w(kN+1), \ldots, w(kN+N-1)]^T$, whereas $\mathbf{H}(k)$ and $\mathbf{H}^{\text{ibi}}(k)$ are $N \times N$ upper and lower triangular matrices with entries $[\mathbf{H}(k)]_{n,m} = h(kN+n;\ n-m)$ and $[\mathbf{H}^{\text{ibi}}(k)]_{n,m} = h(kN+n;\ N+n-m)$ for $n, m = 1, \ldots, N$. The second term on the right-hand side (r.h.s.) of (3) captures the interblock interference (IBI) that emerges due to the channel delay spread. The difference between these channel matrices and those in e.g., [31] is that all the channel taps here are time dependent, and $\mathbf{H}(k)$ as well as $\mathbf{H}^{\text{ibi}}(k)$ are no longer Toeplitz matrices.

In this paper, we wish to design the optimal training input for channel estimation, which amounts to selecting the number of training symbols per block, the placement of training symbols, and the power allocation between training and information symbols, based on conditional mutual information and channel estimation error criteria. Our joint consideration of these criteria is intuitively appealing because of the apparent tradeoff: Using more training symbols of higher power improves channel estimation but also leads to reduced channel capacity.

III. Channel Estimation

Since the channel coefficients $h_q(\lfloor i/N \rfloor; l)$ in (1) are time-invariant over NT_s s, channel estimation has to be performed every N symbols. To enable low-complexity block-by-block processing at the receiver, we need to remove the IBI not only across blocks but also within each block. There are at least two ways to eliminate IBI (see, e.g., [31]): One consists of introducing redundancy at the transmitter through the cyclic-prefix and then discarding those "channel-contaminated" redundant symbols at the receiver, and the other is to put guard zeros per transmitted (sub)block. We adopt the latter in this paper. Hence, we construct $\mathbf{u}(k)$ to satisfy the condition:

C1) Each block $\mathbf{u}(k)$ has the form $[\overline{\mathbf{u}}^T(k)\mathbf{0}_{1\times L}]^T$, where the $(N-L)\times 1$ vector $\overline{\mathbf{u}}(k)$ contains N_s information symbols and $N_b - L \geq 0$ training symbols.

We view the L trailing zeros in $\mathbf{u}(k)$ as part of the training symbols. Since $\mathbf{H}^{\text{ibi}}(k)\mathbf{u}(k-1) = \mathbf{0}$, C1) guarantees the elimination of IBI from block to block. As shown in Fig. 2, the placement of these symbols in $\mathbf{u}(k)$ can be expressed as

$$\mathbf{u}(k) = [\mathbf{s}_1^T(k), \mathbf{b}_1^T(k), \ldots, \mathbf{s}_P^T(k), \mathbf{b}_P^T(k)]^T, \qquad \forall k \quad (4)$$

where we group consecutive information symbols and training symbols in sub-blocks: $\mathbf{s}_p(k)$ and $\mathbf{b}_p(k)$ of lengths $N_{s,p}$ and $N_{b,p}$, respectively. Notice that these parameters satisfy $\sum_{p=1}^P N_{s,p} = N_s$, $\sum_{p=1}^P N_{b,p} = N_b$, and $N_s + N_b = N$. Condition C1) requires one to choose $N_b \geq L$ and the last L entries of $\mathbf{b}_P(k)$ to be zero.

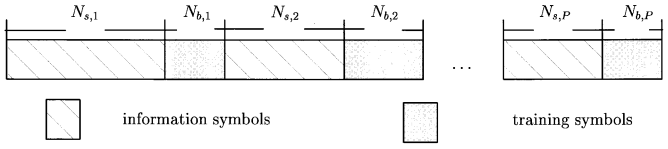


Fig. 2. Structure of the transmitted block $\mathbf{u}(k)$.

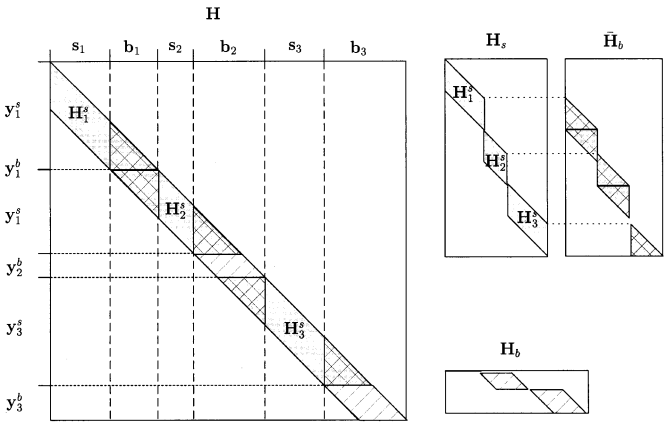


Fig. 3. Partition of the matrix $\mathbf{H}$ in (6).

Taking C1) into account, we rewrite the input–output relationship (3) as

$$\mathbf{y}(k) = \mathbf{H}(k)\mathbf{u}(k) + \mathbf{w}(k). \qquad (5)$$

We wish to estimate $\mathbf{H}(k)$ based on $\mathbf{y}(k)$ and our optimally designed training symbols in $\mathbf{u}(k)$ and then recover the unknown information symbols $\{\mathbf{s}_p(k)\}_{p=1}^P$ based on the estimated $\hat{\mathbf{H}}(k)$. This decoupling of channel from symbol estimation is the motivation behind our separable block structure in (4). It also enables separation of each received block $\mathbf{y}(k)$ into two types of received sub-blocks: one, defined as $\mathbf{y}_b(k)$, that depends only on $\mathbf{H}(k)$ and $\{\mathbf{b}_p(k)\}_{p=1}^P$ and a second, defined as $\mathbf{y}_s(k)$, that depends on $\mathbf{H}(k)$, $\{\mathbf{s}_p(k)\}_{p=1}^P$, and $\{\mathbf{b}_p(k)\}_{p=1}^P$. Because the following analysis for both channel estimation and symbol detection is based on a single block, we omit the block index k and subsequently deal with the input–output relationship [c.f. (1) and (5)]:

$$\mathbf{y} = \mathbf{H}\mathbf{u} + \mathbf{w}, \qquad \text{with } \mathbf{H} := \sum_{q=0}^{Q} \mathbf{D}_q \mathbf{H}_q \qquad (6)$$

where $\mathbf{D}_q := \text{diag}[1, e^{j\omega_q}, \ldots, e^{j\omega_q(N-1)}]$, and $\mathbf{H}_q$ is a lower triangular Toeplitz matrix with first column $[h_q(0), \ldots, h_q(L), 0, \ldots, 0]^T$. Corresponding to the separation of $\mathbf{y}$ to $\mathbf{y}_s$ and $\mathbf{y}_b$, the channel matrix $\mathbf{H}$ can be split into three matrices, namely, $\mathbf{H}_s$, $\mathbf{H}_b$, and $\overline{\mathbf{H}}_b$, which are depicted in Fig. 3. Each of them is constructed from sub-blocks of $\mathbf{H}$. After the separation of $\mathbf{y}$, we have two input–output relationships

$$\mathbf{y}_s = \mathbf{H}_s\mathbf{s} + \overline{\mathbf{H}}_b\overline{\mathbf{b}} + \mathbf{w}_s \qquad (7)$$

$$\mathbf{y}_b = \mathbf{H}_b\mathbf{b} + \mathbf{w}_b \qquad (8)$$

where $\mathbf{s} := [\mathbf{s}_1^T, \ldots, \mathbf{s}_P^T]^T$, and $\mathbf{b} := [\mathbf{b}_1^T, \ldots, \mathbf{b}_P^T]^T$, $\overline{\mathbf{b}}$ contains the first L and the last L entries of $\mathbf{b}_p$, $\forall p$, whereas $\mathbf{w}_s$ and $\mathbf{w}_b$ denote the corresponding noise vectors. The term $\overline{\mathbf{H}}_b\overline{\mathbf{b}}$

captures the interference of the training sub-blocks to their adjacent information sub-blocks.

Focusing first on channel estimation, we start from the training input–output relationship (8). Based on (6) and Fig. 3, $\mathbf{y}_b$ can be written as

$$\mathbf{y}_b := \begin{bmatrix} \mathbf{y}_1^b \\ \vdots \\ \mathbf{y}_P^b \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1^b \mathbf{b}_1 \\ \vdots \\ \mathbf{H}_P^b \mathbf{b}_P \end{bmatrix} + \mathbf{w}_b \tag{9}$$

where $\mathbf{y}_p^b := \mathbf{H}_p^b \mathbf{b}_p + \mathbf{w}_p^b$, and $\forall\, p \in [1,\, P]$, with $\mathbf{H}_p^b$, shown at the bottom of the page, n_p is the index of the first element of $\mathbf{y}_p$ in $\mathbf{y}$, and $\mathbf{w}_b = [(\mathbf{w}_1^b)^T \cdots (\mathbf{w}_P^b)^T]^T$ is the corresponding noise block. It is clear that when $N_{b,\,p} \leq L$, the matrix $\mathbf{H}_p^b$ disappears, and $\mathbf{b}_p$ does not contain sufficient training symbols for channel estimation. Hence, we have the condition shown at the bottom of the page.

C2) The length of each training sub-block $\mathbf{b}_p$ is at least $L+1$, i.e., $N_{b,\,p} \geq L+1, \forall\, p \in [1,\, P]$.

Condition C2) shows that once we insert pilot symbols, we should group them in sub-blocks of size at least $L+1$. Same condition can be found in [11] and [32], which dealt with frequency-selective channels only. Observing the dimensionality of $\mathbf{H}_p^b$, we deduce that out of the N_b pilot symbols transmitted, we receive at most $N_b - PL$ pilot-dependent observations without interference from the unknown information symbols. Since we have $(Q+1)(L+1)$ unknown coefficients [c.f. (1)], to ensure uniqueness in estimating the channel using linear equations, we need the total number of training symbols to satisfy

$$N_b \geq PL + (Q+1)(L+1). \tag{10}$$

Therefore, the minimum number of pilot symbols N_b for estimating doubly selective channels is $L + (Q+1)(L+1)$, when $P = 1$. Selecting $P = 1$ corresponds to the preamble-based training method. From a bandwidth efficiency point of view, this method is optimal. Is this also optimal when we consider mutual information based on *estimated* channels? Recall the tradeoff that emerges: Increasing the number of pilots improves the accuracy of channel estimators, but at the same time, it reduces the rate. In the following, we will answer this question and delineate the tradeoff.

Going back to (9), and based on (1), we can write $\mathbf{H}_p^b := \sum_{q=0}^{Q} \mathbf{D}_{q,\,p}^b \mathbf{H}_{q,\,p}^b$, where $\mathbf{H}_{q,\,p}^b$ and $\mathbf{D}_{q,\,p}^b$ are corresponding sub-matrices from $\mathbf{D}_q$ and $\mathbf{H}_q$ in (6). Plugging $\mathbf{H}_p^b$ into (9), we obtain

$$\mathbf{y}_b = \sum_{q=0}^{Q} \begin{bmatrix} \mathbf{D}_{q,\,1}^b \mathbf{H}_{q,\,1}^b \mathbf{b}_1 \\ \vdots \\ \mathbf{D}_{q,\,P}^b \mathbf{H}_{q,\,P}^b \mathbf{b}_P \end{bmatrix} + \mathbf{w}_b. \tag{11}$$

Due to the commutativity between a Toeplitz (convolution) matrix product with a vector, we have $\mathbf{H}_{q,\,p}^b \mathbf{b}_p = \mathbf{B}_p \mathbf{h}_q$, where $\mathbf{B}_p$ is an $(N_{b,\,p} - L) \times (L+1)$ Toeplitz matrix given by

$$\mathbf{B}_p = \begin{bmatrix} b_{p,\,L} & \cdots & b_{p,\,0} \\ \vdots & \cdots & \vdots \\ b_{p,\,N_{b,\,p}-1} & \cdots & b_{p,\,N_{b,\,p}-L-1} \end{bmatrix}$$

and

$$\mathbf{h}_q := \begin{bmatrix} h_q(0) \\ \vdots \\ h_q(L) \end{bmatrix}$$

with $b_{p,\,n}$ denoting the $(n+1)$st entry of $\mathbf{b}_p$. Hence, the input–output relationship in (11) becomes

$$\mathbf{y}_b = \mathbf{\Phi}_b \mathbf{h} + \mathbf{w}_b \tag{12}$$

where

$$\Phi_b := \begin{bmatrix} \mathbf{D}_{0,\,1}^b \mathbf{B}_1 & \cdots & \mathbf{D}_{Q,\,1}^b \mathbf{B}_1 \\ \vdots & \cdots & \vdots \\ \mathbf{D}_{0,\,P}^b \mathbf{B}_P & \cdots & \mathbf{D}_{Q,\,P}^b \mathbf{B}_P \end{bmatrix} \tag{13}$$

and

$$\mathbf{h} := [\,\mathbf{h}_0^T \quad \cdots \quad \mathbf{h}_Q^T\,]^T. \tag{14}$$

Similar to [21], we will rely on the Wiener solution of (12) that yields the linear[1] MMSE (LMMSE) channel estimator

$$\hat{\mathbf{h}} = \frac{1}{\sigma_w^2} \left(\mathbf{R}_h^{-1} + \frac{1}{\sigma_w^2} \mathbf{\Phi}_b^{\mathcal{H}} \mathbf{\Phi}_b \right)^{-1} \mathbf{\Phi}_b^{\mathcal{H}} \mathbf{y}_b \tag{15}$$

which requires $\mathbf{R}_h := E[\mathbf{h}\mathbf{h}^{\mathcal{H}}]$ to be known at the receiver.

Defining the channel error as $\tilde{\mathbf{h}} := \mathbf{h} - \hat{\mathbf{h}}$, we can express its correlation as

$$\mathbf{R}_{\tilde{h}} := E[\tilde{\mathbf{h}}\tilde{\mathbf{h}}^{\mathcal{H}}] = \left(\mathbf{R}_h^{-1} + \frac{1}{\sigma_w^2} \mathbf{\Phi}_b^{\mathcal{H}} \mathbf{\Phi}_b \right)^{-1} \tag{16}$$

and the mean square error of $\hat{\mathbf{h}}$ as

$$\sigma_{\tilde{h}}^2 := \mathrm{tr}(\mathbf{R}_{\tilde{h}}) = \mathrm{tr} \left(\left(\mathbf{R}_h^{-1} + \frac{1}{\sigma_w^2} \mathbf{\Phi}_b^{\mathcal{H}} \mathbf{\Phi}_b \right)^{-1} \right). \tag{17}$$

It is clear from (13) that the placement of training symbols affects Φ_b and, consequently, $\sigma_{\tilde{h}}^2$. To facilitate our subsequent analysis, we suppose the following.

A3) The channel coefficients $h_q(l)$ are independent, i.e., $\mathbf{R}_h$ is a diagonal matrix with trace $\mathrm{tr}(\mathbf{R}_h) = 1$.

[1]Under A2), $\mathbf{h}$ is Gaussian in the linear model (12); hence, the LMMSE coincides with MMSE optimal channel estimator.

$$\mathbf{H}_p^b = \begin{bmatrix} h(n_p;\, L) & \cdots & h(n_p;\, 0) & \cdots & 0 \\ & \ddots & & \ddots & \\ 0 & \cdots & h(n_p + N_{b,\,p} - L - 1;\, L) & \cdots & h(n_p + N_{b,\,p} - L - 1;\, 0) \end{bmatrix}_{(N_{b,\,p}-L) \times N_{b,\,p}}.$$

Note that A3) will not affect the optimality of our training design, simply because no CSI is assumed available at the transmitter.

Using [21, Lemma 7] and A3), it can be shown that $\sigma_{\tilde{h}}^2$ in (17) is lower bounded as follows:

$$\mathrm{tr}\left(\left(\mathbf{R}_h^{-1}+\frac{1}{\sigma_w^2}\,\mathbf{\Phi}_b^{\mathcal{H}}\mathbf{\Phi}_b\right)^{-1}\right)\geq\sum_m\frac{1}{[\mathbf{R}_h^{-1}+\frac{1}{\sigma_w^2}\,\mathbf{\Phi}_b^{\mathcal{H}}\mathbf{\Phi}_b]_{m,m}} \tag{18}$$

where the equality holds if and only if $\Phi_b^{\mathcal{H}}\Phi_b$ is a diagonal matrix. Therefore, the following condition is required for our training strategy to attain the channel MMSE:

C3) For fixed N_b and N_s, the training symbols should be inserted so that the matrix $\Phi_b^{\mathcal{H}}\Phi_b$ is diagonal.

Condition C3) coincides with that in [1], [9], [11], and [32].

Although we have set our channel estimate in (15), and we have built C1)–C3), there are additional training parameters that have not been decided, such as the placement and the optimal number of training symbols. These parameters affect the performance of the channel estimator in (15), the effective transmission rate ($\eta = N_s/N$), the mutual information, as well as the bit error rate (BER). In the following, we will select these training parameters by optimizing an average capacity bound. Prior to this, however, we will show that optimizing this average capacity bound also minimizes the channel MMSE.

IV. Linking Capacity With Channel Estimation

It is not easy to evaluate the average capacity of an unknown random channel that has to be estimated. Instead, we will derive an upper bound and a lower bound. To design our optimal training parameters, we will maximize the capacity lower bound, and view the upper bound as a benchmark for the maximum achievable rate.

Let $\mathcal{P}$ denote the total transmit-power per block, $\mathcal{P}_s$ the power allocated to the information signal part, and $\mathcal{P}_b$ the power assigned to the training part. Before we consider optimal power allocation, we suppose that $\mathcal{P}_s$ and $\mathcal{P}_b$ are fixed. Let $\hat{\mathbf{H}}$ be any estimator of $\mathbf{H}$ in (6). Since training symbols $\mathbf{b}$ do not convey information, for a fixed power $\mathcal{P}_s := E[||\mathbf{s}||^2]$, the conditional mutual information between transmitted information symbols and received symbols in (7) is denoted as $\mathcal{I}(\mathbf{y}_s;\ \mathbf{s}|\hat{\mathbf{h}})$ for each realization of $\mathbf{H}$. The channel capacity averaged over the random channel $\mathbf{H}$ is defined as

$$C := \frac{1}{N}E\left[\max_{p_s(\cdot),\,\mathcal{P}_s:=E[||\mathbf{s}||^2]}\mathcal{I}(\mathbf{y}_s;\ \mathbf{s}|\hat{\mathbf{h}})\right]\ \text{bits/s/Hz} \tag{19}$$

where $p_s(\cdot)$ denotes the probability density function of $\mathbf{s}$.

A. Upper Bound on Capacity With Perfectly Known Channel

Suppose first that the channel estimation is perfect, i.e., $\hat{\mathbf{H}} \equiv \mathbf{H}$. Similar to (19), the average capacity in this ideal case is defined as

$$\overline{C} := \frac{1}{N}E\left[\max_{p_s(\cdot),\,\mathcal{P}_s:=E[||\mathbf{s}||^2]}\mathcal{I}(\mathbf{y}_s;\ \mathbf{s}|\mathbf{H})\right]\ \text{bits/s/Hz}. \tag{20}$$

From (7), we know that $\mathbf{y}_s = \mathbf{H}_s\mathbf{s} + \overline{\mathbf{H}}_b\overline{\mathbf{b}} + \mathbf{w}_s$, where $\overline{\mathbf{H}}_b$ is the corresponding channel matrix for $\overline{\mathbf{b}}$ (see also Fig. 3 for the structures of $\mathbf{H}_s$ and $\overline{\mathbf{H}}_b$). Because $\overline{\mathbf{H}}_b$ and $\overline{\mathbf{b}}$ in (7) are known in the ideal case, by defining $\mathbf{y}_s' := \mathbf{y}_s - \overline{\mathbf{H}}_b\overline{\mathbf{b}}$, it can be verified that $\mathcal{I}(\mathbf{y}_s;\mathbf{s}|\mathbf{h}) = \mathcal{I}(\mathbf{y}_s';\mathbf{s}|\mathbf{h})$. To maximize $\mathcal{I}(\mathbf{y}_s';\ \mathbf{s}|\mathbf{H})$, we establish the following lemma.

Lemma 1: If the information bearing block $\mathbf{s}$ is Gaussian distributed, then the mutual information $\mathcal{I}(\mathbf{y}_s';\ \mathbf{s}|\mathbf{h})$ is maximized. Furthermore, the capacity upper bound $\overline{C}$ in (20) can be expressed as

$$\overline{C} := \frac{1}{N}E\left[\max_{\mathbf{R}_s,\,\mathcal{P}_s:=E[||\mathbf{s}||^2]}\log\det\cdot\left(\mathbf{I}_{N_s+LP}+\frac{1}{\sigma_w^2}\,\mathbf{H}_s\mathbf{R}_s\mathbf{H}_s^{\mathcal{H}}\right)\right]\ \text{bits/s/Hz}. \tag{21}$$

Proof: See Appendix A. ∎

Although $\mathbf{s}$ is generally non-Gaussian, if N_s is sufficiently large and $\mathbf{s}$ is channel coded (or linearly precoded as in [31]), then $\mathbf{s}$ will be approximately Gaussian. Thus, in the following, we assume the following.

A4) The information-bearing symbol block $\mathbf{s}$ is zero-mean Gaussian with covariance $\mathbf{R}_s = \overline{\mathcal{P}}_s\mathbf{I}_{N_s}$, and $\overline{\mathcal{P}}_s := \mathcal{P}_s/N_s$.

Here, we select $\mathbf{R}_s = \overline{\mathcal{P}}_s\mathbf{I}_{N_s}$ because there is no CSI at the transmitter and, hence, nonuniform power-loading has no basis. We underscore that $\overline{C}$ is an upper bound on the average channel capacity with estimated channels because it expresses the ideal channel capacity without channel estimation error.

B. Lower Bound on Capacity With LMMSE Channel Estimation

Consider now that the estimate of $\mathbf{H}$ is imperfect. Define $\hat{\mathbf{H}}_s$ as the estimate of $\mathbf{H}_s$ and $\hat{\overline{\mathbf{H}}}_b$ as the estimate of $\overline{\mathbf{H}}_b$. Since $\overline{\mathbf{b}}$ and $\hat{\overline{\mathbf{H}}}_b$ are known, we subtract $\hat{\mathbf{H}}_b\overline{\mathbf{b}}$ from $\mathbf{y}_s$. Thus, we have [c.f. (7)]

$$\mathbf{y}_s' := \mathbf{y}_s - \hat{\overline{\mathbf{H}}}_b\overline{\mathbf{b}} = \hat{\mathbf{H}}_s\mathbf{s} + (\mathbf{H}_s - \hat{\mathbf{H}}_s)\mathbf{s} + (\overline{\mathbf{H}}_b - \hat{\overline{\mathbf{H}}}_b)\overline{\mathbf{b}} + \mathbf{w}_s. \tag{22}$$

Using (7) and (22), it is easy to verify that $\mathcal{I}(\mathbf{y}_s';\ \mathbf{s}|\hat{\mathbf{H}}_s) = \mathcal{I}(\mathbf{y}_s;\ \mathbf{s}|\hat{\mathbf{H}}_s)$. Define $\tilde{\mathbf{H}}_s := \mathbf{H}_s - \hat{\mathbf{H}}_s$, $\tilde{\mathbf{H}}_b := \overline{\mathbf{H}}_b - \hat{\overline{\mathbf{H}}}_b$, and $\mathbf{v} := \tilde{\mathbf{H}}_s\mathbf{s} + \tilde{\overline{\mathbf{H}}}_b\overline{\mathbf{b}} + \mathbf{w}_s$. In general, $\mathbf{v}$ is non-Gaussian distributed with correlation matrix $\mathbf{R}_v := E[\mathbf{v}\mathbf{v}^{\mathcal{H}}]$ given by

$$\mathbf{R}_v = \overline{\mathcal{P}}_s E[\tilde{\mathbf{H}}_s\tilde{\mathbf{H}}_s^{\mathcal{H}}] + E[\tilde{\mathbf{H}}_b\overline{\mathbf{b}}\overline{\mathbf{b}}^{\mathcal{H}}\tilde{\mathbf{H}}_b^{\mathcal{H}}] + \sigma_w^2\mathbf{I}_{N_s+LP} \tag{23}$$

where $E[\tilde{\mathbf{H}}_s\mathbf{s}\overline{\mathbf{b}}\tilde{\mathbf{H}}_s^{\mathcal{H}}] = \mathbf{0}$ because of A4). Because of the non-Gaussianity of $\mathbf{v}$, it is not easy to obtain a closed form of the average capacity. In the following, we propose a lower bound of C in (19).

Lemma 2: When the information-bearing block, $\mathbf{s}$ is Gaussian distributed with fixed power $\mathcal{P}_s$, the average capacity C in (19) is lower-bounded as:

$$C \geq \frac{1}{N}E\left[\max_{\mathbf{R}_s}\log\det\cdot\left(\mathbf{I}_{N_s+LP}+\mathbf{R}_v^{-1}\hat{\mathbf{H}}_s\mathbf{R}_s\hat{\mathbf{H}}_s^{\mathcal{H}}\right)\right]\ \text{bits/s/Hz}. \tag{24}$$

Proof: See Appendix B. ∎

Similar to [1] and [30], we will introduce next a lower bound that is looser than the right-hand side of (24) but easier to handle. Plugging $\mathbf{R}_s = \overline{\mathcal{P}}_s \mathbf{I}_{N_s}$ from A4) into (24), we obtain

$$C \geq \frac{1}{N} E[\log\det\left(\mathbf{I}_{N_s+LP} + \overline{\mathcal{P}}_s \mathbf{R}_v^{-1} \hat{\mathbf{H}}_s \hat{\mathbf{H}}_s^{\mathcal{H}}\right)] := \underline{C}. \quad (25)$$

The right-hand side of (25) offers a lower bound on the average capacity of doubly selective channels. Our objective is to select training parameters so that $\underline{C}$ in (25) is maximized. Certainly, the optimal training parameters should improve both the channel estimator and the associated minimum mean-square error (MMSE) $\sigma_{\tilde{h}}^2$. Interestingly, $\underline{C}$ and the channel MMSE $\sigma_{\tilde{h}}^2$ are linked. To establish this link, we will introduce two useful lemmas.

Lemma 3: Suppose C1)–C3) and A1)–A4) hold true and that the information symbol power $\overline{\mathcal{P}}_s$ and the sub-block lengths $N_{b,p}$ and $N_{s,p}$ are fixed. Then, maximizing $\underline{C}$ in (25) is equivalent to minimizing $\mathbf{R}_v$ in (23), at high signal-to-noise ratio (SNR).

Proof: See Appendix C. ∎

Although $\mathbf{R}_v$ in (23) depends on $\sigma_{\tilde{h}}^2$, this dependence is not explicit. The following lemma provides an explicit relationship between the two.

Lemma 4: Consider a fixed number of training symbols N_b adhering to C1) and C2). Among all $\mathbf{b}_p$ choices that satisfy C3) and lead to identical $\mathbf{R}_{\tilde{h}}$, the design that satisfies $N_{b,p} \geq 2L+1$ and has the first L and the last L entries of $\mathbf{b}_p, \forall\, p \in [1,\, P]$ equal to zero achieves the minimum $\mathbf{R}_v$.

Proof: See Appendix D. ∎

Based on Lemmas 3 and 4, we modify condition C2) to the following.

C2′) The training sub-block is $\mathbf{b}_p := [\mathbf{0}_L^T \overline{\mathbf{b}}_p^T \mathbf{0}_L^T]^T, \forall\, p \in [1,\, P]$, with the length of $\overline{\mathbf{b}}_p$, $N_{\overline{b},p} \geq 1$.

Notice that the L zeros between the information sub-blocks $\mathbf{s}_p$ and the training sub-blocks $\overline{\mathbf{b}}_p$ eliminate the intersub-block interference. Condition C2′) implies that $N_{b,p} \geq 2L+1$. Based on the assumptions and design conditions we have introduced so far, we are ready to establish the link between channel MMSE in (18) and the lower bound $\underline{C}$ in (25).

Proposition 1: Suppose A1)–A4) and C1)–C3) hold true. If $N_{s,p} \gg 2L, \forall\, p$, then for fixed $N_{s,p}$ and $N_{b,p}$, the minimization of $\sigma_{\tilde{h}}^2$ in (18) is equivalent to the maximization of $\underline{C}$ in (25).

Proof: Defining $\psi_{q,l} := E[\tilde{h}_q(l)\tilde{h}_q^*(l)]$ and relying on C3), we can express the correlation matrix of $\tilde{\mathbf{h}}$ in (16) as

$$\mathbf{R}_{\tilde{h}} = \text{diag}[\psi_{0,0},\, \ldots,\, \psi_{Q,L}]. \quad (26)$$

Since $\mathbf{D}_q$ is known, $\tilde{\mathbf{H}}_s := \mathbf{H}_s - \hat{\mathbf{H}}_s$ is a block-diagonal matrix [c.f. Fig. 3], and because $E[\tilde{h}_{q_1}(l_1)\tilde{h}_{q_2}^*(l_2)] = 0, \forall\, l_1 \neq l_2$ or $\forall\, q_1 \neq q_2$, the correlation matrix of $\tilde{\mathbf{H}}_s$ can be written as

$$E[\tilde{\mathbf{H}}_s \tilde{\mathbf{H}}_s^{\mathcal{H}}] = \sum_{q=0}^{Q} \mathbf{D}_q E[\tilde{\mathbf{H}}_q^s (\tilde{\mathbf{H}}_q^s)^{\mathcal{H}}] \mathbf{D}_q^{\mathcal{H}} \quad (27)$$

where

$$\tilde{\mathbf{H}}_q^s = \begin{bmatrix} \tilde{\mathbf{H}}_{q,1}^s & & \\ & \ddots & \\ & & \tilde{\mathbf{H}}_{q,P}^s \end{bmatrix}$$

and $\tilde{\mathbf{H}}_{q,p}^s$ is a lower triangular Toeplitz matrix with first column $[\tilde{h}_q(0),\, \ldots,\, \tilde{h}_q(L),\, 0,\, \ldots,\, 0]^T$. From (26), we can detail (27) as

$$E[\tilde{\mathbf{H}}_{q,p}^s (\tilde{\mathbf{H}}_{q,p}^s)^{\mathcal{H}}] = \text{diag}\left[\psi_{q,0} \sum_{l=0}^{1} \psi_{q,l} \cdots \sum_{l=0}^{L} \psi_{q,l} \cdots \sum_{l=0}^{L} \psi_{q,l} \cdots \psi_{q,L}\right]. \quad (28)$$

Equation (28) shows that the correlation matrix of $\tilde{\mathbf{H}}_q^s$ (and, thus, $\tilde{\mathbf{H}}_s$) is a diagonal matrix. In addition, we notice that selecting $N_{s,p} \gg 2L$ allows one to approximate the correlation matrix of $\tilde{\mathbf{H}}_s$ as follows [c.f. (28)]:

$$E[\tilde{\mathbf{H}}_s \tilde{\mathbf{H}}_s^{\mathcal{H}}] \approx \sum_{q=0}^{Q} \sum_{l=0}^{L} \psi_{q,l} \mathbf{I}_{N_s+PL} = \sigma_{\tilde{h}}^2 \mathbf{I}_{N_s+PL}.$$

Considering C2′) and C3), we can write the correlation matrix $\mathbf{R}_v$ in (23) as

$$\mathbf{R}_v \approx \sigma_{\tilde{h}}^2 \overline{\mathcal{P}}_s \mathbf{I}_{N_s+PL} + \sigma_w^2 \mathbf{I}_{N_s+PL}. \quad (29)$$

We deduce from (29) that as $\sigma_{\tilde{h}}^2$ decreases, $\mathbf{R}_v$ decreases, and from Lemma 3, we infer that $\underline{C}$ increases accordingly, i.e., better channel estimation implies higher average capacity. ∎

V. Designing Optimal Training Parameters

In the previous section, we linked the LMMSE channel estimation with the maximum lower bound of the average channel capacity. In this section, we will capitalize on this link to design our optimal training parameters. Specifically, we will answer the following basic questions: How should we place the training symbols? How many training symbols should be inserted per block? How much power should be allocated to the training symbols?

A. Optimal Placement of Pilots

Since we have adopted the LMMSE channel estimator, we start from (15)–(17). In (17), we expressed the MMSE channel estimation error as $\sigma_{\tilde{h}}^2 = \text{tr}\left(\left(\mathbf{R}_h^{-1} + (1/\sigma_w^2)\Phi_b^{\mathcal{H}}\Phi_b\right)^{-1}\right)$. Now, we will design Φ_b [which certainly depends on $\mathbf{B}_p$ as per (13)] so that $\sigma_{\tilde{h}}^2$ is minimized subject to the power constraint on the total power of pilots. Under C3), the right-hand side of (18) satisfies

$$\sigma_{\tilde{h}}^2 = \sum_m \frac{1}{[\mathbf{R}_h^{-1} + \frac{1}{\sigma_w^2}\mathbf{\Phi}_b^{\mathcal{H}}\mathbf{\Phi}_b]_{m,m}} \geq \sum_m \frac{1}{[\mathbf{R}_h^{-1} + \frac{\mathcal{P}_b}{\sigma_w^2}\mathbf{I}]_{m,m}} \quad (30)$$

where the second equality holds if and only if $\Phi_b^{\mathcal{H}}\Phi_b = \mathcal{P}_b\mathbf{I}$. Based on the structure of Φ_b, we infer that equivalently, we need two conditions to be fulfilled [c.f. (13)]:

$$\sum_{p=1}^{P} \mathbf{B}_p^{\mathcal{H}} \mathbf{B}_p = \mathcal{P}_b \mathbf{I}_{N_{b,p}} \quad (31)$$

$$\sum_{p=1}^{P} \mathbf{B}_p^{\mathcal{H}} \mathbf{D}_{q_1,p}^{\mathcal{H}} \mathbf{D}_{q_2,p} \mathbf{B}_p = \mathbf{0}, \qquad \forall\, q_1 \neq q_2. \quad (32)$$

A general placement satisfying (31) and (32) is not easy to obtain directly. As a first step, we will need the following lemma to gain further insight on the optimal placement.

Lemma 5: Consider a fixed number of training symbols N_b, information symbols $N_s > 2L$, power $\mathcal{P}_s$, and a number of sub-blocks P per block. If N_s is an integer multiple of P, then equally long information sub-blocks maximize the lower bound of capacity $\underline{C}$. The length of the information sub-blocks is $\overline{N}_s := N_s/P$.

Proof: See Appendix E. ∎

The following proposition provides sufficient conditions to achieve (31) and (32).

Proposition 2: Suppose A1)–A4) hold true. For fixed $\mathcal{P}_s$ and $\mathcal{P}_b$, the following placement is optimal: All information sub-blocks have identical block lengths, i.e., $N_{s,p} = \overline{N}_s, \forall p$; the pilot sub-blocks have identical structure $[\mathbf{0}_L^T\ b\ \mathbf{0}_L^T]^T, \forall p$, and they are equipowered with $b = \overline{\mathcal{P}}_b := \mathcal{P}_b/P$.

Proof: First, we will confirm that conditions C1)–C3) hold true. Then, according to Proposition 1, we will verify that $\underline{C}$ is maximized, and finally, we will check whether $\sigma^2_{\tilde{h}}$ is also minimized.

If $\forall p$, $N_{\overline{b},p} = 1$, and $\mathbf{B}_p = \sqrt{\overline{\mathcal{P}}_b}\mathbf{I}_{L+1}$, then we have that $\mathbf{B}_p^{\mathcal{H}}\mathbf{B}_p = \overline{\mathcal{P}}_b\mathbf{I}_{L+1}$. Therefore, $\sum_{p=1}^{P}\mathbf{B}_p^{\mathcal{H}}\mathbf{B}_p = \mathcal{P}_b\mathbf{I}_{L+1}$ is a diagonal matrix. Thus, we have checked the first condition in (31). Plugging $\mathbf{B}_p$ into the left-hand side of (32), we obtain

$$\sum_{p=1}^{P}\mathbf{B}_p^{\mathcal{H}}\mathbf{D}_{q_1,p}^{\mathcal{H}}\mathbf{D}_{q_2,p}\mathbf{B}_p = \overline{\mathcal{P}}_b\sum_{p=1}^{P}\overline{\mathbf{D}}_{q_1,p}^{\mathcal{H}}\overline{\mathbf{D}}_{q_2,p}$$

where $\overline{\mathbf{D}}_{q,p}$ contains the first $L+1$ columns and the first $L+1$ rows of $\mathbf{D}_{q,p}$. Because the BEM frequencies are equispaced, it follows that $N_{s,p} = \overline{N}_s$ and that $N_{b,p} = 2L+1$. By defining the difference between two consecutive BEM frequencies as $\omega_q - \omega_{q-1} = 2\pi/N$, we find that

$$\overline{\mathbf{D}}_{q_1,p}^{\mathcal{H}}\overline{\mathbf{D}}_{q_2,p} = e^{j(2\pi/N)(q_2-q_1)(\overline{N}_s+2L+1)(p-1)} \cdot\mathrm{diag}\left[1,\ e^{j(2\pi/N)(q_2-q_1)},\ \ldots,\ e^{j(2\pi/N)(q_2-q_1)(\overline{N}_s+L)}\right].$$

Now, we know that the transmitted block length should be $N = (\overline{N}_s + 2L + 1)P$; hence

$$\sum_{p=1}^{P}\overline{\mathbf{D}}_{q_1,p}^{\mathcal{H}}\overline{\mathbf{D}}_{q_2,p} = \begin{cases} P\mathbf{I}_{L+1}, & q_1 = q_2 \\ \mathbf{0}, & q_1 \neq q_2 \end{cases}$$

which implies that the proposed placement satisfies (32), as well as C3)

$$\mathbf{\Phi}_b^{\mathcal{H}}\mathbf{\Phi}_b = \mathcal{P}_b\mathbf{I}_{(Q+1)(L+1)}. \tag{33}$$

The MMSE $\sigma^2_{\tilde{h}}$ in (30) has thus been achieved, and $\underline{C}$ has been maximized per Proposition 1. ∎

Guided by Proposition 2, we can now finalize the structure of our transmitted block $\mathbf{u}$ as

$$\mathbf{u} = [\mathbf{s}_1^T\ \mathbf{0}_L^T\ b\ \mathbf{0}_L^T\cdots\mathbf{s}_P^T\ \mathbf{0}_L^T\ b\ \mathbf{0}_L^T]^T, \qquad b = \sqrt{\overline{\mathcal{P}}_b}. \tag{34}$$

From Proposition 2, we have obtained that $N_{b,p} = 2L+1$. To satisfy the condition in (10), we have that $P \geq Q+1$. To complete the optimality claims on the number of training symbols per sub-block, we establish the following proposition.

Proposition 3: If the powers $\mathcal{P}_s$ and $\mathcal{P}_b$ are fixed, the number of sub-blocks $P \geq Q+1$, and $N_{b,p} \geq 2L+1$, then as $N_{b,p}\ \forall p$ and/or P increase, $\underline{C}$ decreases.

Notice that when $N_{b,p} < 2L+1$ and $P < Q+1$, the minimum $\mathbf{R}_v$ in (30) cannot be guaranteed, per Lemma 4.

B. Channel MMSE and Capacity With Optimal Placement

We have seen that (34) offers the optimal placement of pilot and information symbols per block that not only maximizes $\underline{C}$ but at the same time minimizes the LMMSE channel estimation error. Under the Gaussian channel assumption, the latter coincides with the channel MMSE, and thus, it benchmarks estimation performance when $\mathbf{R}_h$ is known at the receiver. In this subsection, we will derive this benchmark channel MMSE for the optimal placement when $\mathbf{R}_h$ is known, as well as when $\mathbf{R}_h$ is unknown. Furthermore, we will develop in closed form the maximum $\underline{C}$ when the optimum placement of (34) is used. This is practically important because it allows one to predict the optimal average-rate possible through doubly selective fading channels when optimal training is adopted for channel estimation purposes.

If the channel coefficients are independent (but not necessarily identically distributed), then plugging (33) into (16), we can explicitly write $\mathbf{R}_{\tilde{h}}$ as

$$\mathbf{R}_{\tilde{h}} = \mathrm{diag}\left[\frac{\sigma^2_{0,0}\sigma^2_w}{\sigma^2_w + \mathcal{P}_b\sigma^2_{0,0}}\cdots\frac{\sigma^2_{Q,L}\sigma^2_w}{\sigma^2_w + \mathcal{P}_b\sigma^2_{Q,L}}\right] \tag{35}$$

where $\sigma^2_{q,l}$ is the variance of $h_q(l)$. The $\mathrm{tr}(\mathbf{R}_{\tilde{h}})$ benchmarks the performance of our doubly selective channel estimator when the $h_q(l)$s are independent with known variances.

For channel-estimation purposes, the training sequence in (34) is optimal. We note that for a fixed $\mathcal{P}_b$, when $N_{b,p} = 2L+1$, the optimal $\mathrm{tr}(\mathbf{R}_{\tilde{h}})$ will not decrease as long as $P \geq Q+1$ since the lower bound of $\mathrm{tr}(\mathbf{R}_{\tilde{h}})$ in (30) holds for any P. On the other hand, as $\mathcal{P}_b$ increases, $\mathbf{R}_{\tilde{h}}$ will decrease monotonically. However, from a mutual information point of view, this is not the end. Since $\mathcal{P}$ is fixed, if we put more power on training, less power is left for the information symbols to deal with the AWGN. Furthermore, as P increases, the bandwidth efficiency decreases. In the following sub-sections, we will pursue the optimal design for these two parameters.

First, however, we would like to summarize the new conditions implied by Proposition 2 and rewrite $\underline{C}$ based on these conditions.

C4) Select the block length N as a multiple of P, and design each block $\mathbf{u}$ according to (34).

Using (35), we can simplify the correlation matrix of $\tilde{\mathbf{H}}_s$. Since $N_{s,p} = \overline{N}_s$, we can verify that $E[\tilde{\mathbf{H}}^s_{q,p}(\tilde{\mathbf{H}}^s_{q,p})^{\mathcal{H}}]$ does not depend on the index p. Defining $\Psi_q := E[\tilde{\mathbf{H}}^s_{q,p}(\tilde{\mathbf{H}}^s_{q,p})^{\mathcal{H}}]$, it follows that

$$E[\tilde{\mathbf{H}}^s_q(\tilde{\mathbf{H}}^s_q)^{\mathcal{H}}] = \mathbf{I}_P\otimes\mathbf{\Psi}_q$$
$$E[\tilde{\mathbf{H}}_s\tilde{\mathbf{H}}_s^{\mathcal{H}}] = \mathbf{I}_P\otimes\sum_{q=0}^{Q}\mathbf{\Psi}_q.$$

Thanks to the guard zeros surrounding each training symbol in (34), we have that $\overline{\mathbf{b}} = \mathbf{0}$. Hence

$$\mathbf{R}_v = \sigma_w^2\mathbf{I} + \overline{\mathcal{P}}_s\left(\mathbf{I}_P \otimes \sum_{q=0}^{Q}\mathbf{\Psi}_q\right). \tag{36}$$

Using C4), we have that

$$E[\hat{\mathbf{h}}\hat{\mathbf{h}}^{\mathcal{H}}] = \text{diag}\left[\frac{\mathcal{P}_b\sigma_{0,0}^4}{\sigma_w^2+\mathcal{P}_b\sigma_{0,0}^2},\ldots,\frac{\mathcal{P}_b\sigma_{Q,L}^4}{\sigma_w^2+\mathcal{P}_b\sigma_{Q,L}^2}\right] := \text{diag}\left[\varphi_{0,0}\cdots\varphi_{Q,L}\right] \tag{37}$$

and from (37), we find that

$$E[\hat{\mathbf{H}}_{q,p}^s(\hat{\mathbf{H}}_{q,p}^s)^{\mathcal{H}}] = \text{diag}\left[\varphi_{q,0}\sum_{l=0}^{1}\varphi_{q,l}\cdots\sum_{l=0}^{L}\varphi_{q,l}\cdots\sum_{l=0}^{L}\varphi_{q,l}\cdots\varphi_{q,L}\right]. \tag{38}$$

Since $E[\hat{\mathbf{H}}_p^s(\hat{\mathbf{H}}_p^s)^{\mathcal{H}}] = \sum_{q=0}^{Q}E[\hat{\mathbf{H}}_{q,p}^s(\hat{\mathbf{H}}_{q,p}^s)^{\mathcal{H}}]$, we can obtain the normalization factor for $E[\hat{\mathbf{H}}_p^s(\hat{\mathbf{H}}_p^s)^{\mathcal{H}}]$ as

$$\sigma_{\hat{H}}^2 := \text{tr}\left(E[\hat{\mathbf{H}}_p^s(\hat{\mathbf{H}}_p^s)^{\mathcal{H}}]\right) = \overline{N}_s\sum_{q=0}^{Q}\sum_{l=0}^{L}\varphi_{q,l}. \tag{39}$$

Then, we can define the normalized channel matrix as

$$\hat{\mathbf{H}}_p^s = \sigma_{\hat{H}}\hat{\overline{\mathbf{H}}}_p^s, \qquad \forall p.$$

Finally, we deduce that the lower bound on average channel capacity is

$$\underline{C} = \frac{1}{N}\sum_{p=1}^{P}E\left[\log\det\left(\mathbf{I}+\overline{\mathcal{P}}_s\left(\overline{\mathcal{P}}_s\sum_{q=0}^{Q}\mathbf{\Psi}_q+\sigma_w^2\mathbf{I}\right)^{-1}\cdot\sigma_{\hat{H}}^2\hat{\overline{\mathbf{H}}}_p^s(\hat{\overline{\mathbf{H}}}_p^s)^{\mathcal{H}}\right)\right] \text{ bits/s/Hz}. \tag{40}$$

Equation (40) is useful because it relates the lower bound $\underline{C}$ with the number of sub-blocks P and the signal power $\overline{\mathcal{P}}_s$, which in turn depend on the spacing of pilots and the chosen power allocation.

Relying on (37) and (38), we can readily verify the following lemma.

Lemma 6: If A2) holds true, then all $\hat{\overline{\mathbf{H}}}_p^s$ have identical distribution $\forall p \in [1, P]$.

Proof: See Appendix F. ■

Based on Lemma 6, we can rewrite the lower bound on the average capacity as

$$\underline{C} = \frac{P}{N}E\left[\log\det\left(\mathbf{I}_{\overline{N}_s+L}+\sigma_{\hat{H}}^2\overline{\mathcal{P}}_s\left(\overline{\mathcal{P}}_s\sum_{q=0}^{Q}\mathbf{\Psi}_q+\sigma_w^2\mathbf{I}\right)^{-1}\cdot\hat{\overline{\mathbf{H}}}^s(\hat{\overline{\mathbf{H}}}^s)^{\mathcal{H}}\right)\right] \text{ bits/s/Hz}$$

where we used $\hat{\overline{\mathbf{H}}}^s$ to denote $\hat{\overline{\mathbf{H}}}_p^s\ \forall p$. Let us now consider the eigen-decomposition $(\hat{\overline{\mathbf{H}}}^s)^{\mathcal{H}}\hat{\overline{\mathbf{H}}}^s = \mathbf{U}\mathbf{\Lambda}_H\mathbf{U}^{\mathcal{H}}$, where $\mathbf{\Lambda}_H := \text{diag}[\lambda_1, \ldots, \lambda_{\overline{N}_k}]$ is an $\overline{N}_s\times\overline{N}_s$ diagonal matrix with eigenvalues of $(\hat{\overline{\mathbf{H}}}^s)^{\mathcal{H}}\hat{\overline{\mathbf{H}}}^s$ on its main diagonal, and $\mathbf{U}$ is a unitary matrix that contains the corresponding eigen-vectors. In Proposition 1, we have shown that selecting $N \gg 2L$ yields $\mathbf{R}_v \approx (\sigma_{\tilde{h}}^2\overline{\mathcal{P}}_s+\sigma_w^2)\mathbf{I}$. Hence, we have

$$\begin{aligned}\underline{C} &\approx \frac{P}{N}E\left[\log\det\left(\mathbf{I}_{\overline{N}_s}+\frac{\overline{\mathcal{P}}_s\sigma_{\hat{H}}^2}{\sigma_{\tilde{h}}^2\overline{\mathcal{P}}_s+\sigma_w^2}\mathbf{U}\mathbf{\Lambda}_H\mathbf{U}^{\mathcal{H}}\right)\right]\\ &= \frac{P}{N}\sum_{k=1}^{\overline{N}_s}E\left[\log\left(1+\frac{\overline{\mathcal{P}}_s\sigma_{\hat{H}}^2}{\sigma_{\tilde{h}}^2\overline{\mathcal{P}}_s+\sigma_w^2}\lambda_k\right)\right] \text{ bits/s/Hz}\end{aligned} \tag{41}$$

where in deriving (41), we used the fact that $\det(\mathbf{I}+\mathbf{AB}) = \det(\mathbf{I}+\mathbf{BA})$ for matrices $\mathbf{A}$ and $\mathbf{B}$ with matching dimensions. Equation (41) is similar to that derived in [21]. The key difference here is that the λ_ks are not identically distributed, in general. This will lead to a looser lower bound on the average capacity. Let the effective SNR be defined as

$$\rho_{\text{eff}} = \frac{\overline{\mathcal{P}}_s\sigma_{\hat{H}}^2}{\sigma_w^2+\overline{\mathcal{P}}_s\sigma_{\tilde{h}}^2}. \tag{42}$$

Since $\overline{N}_sP = N - P(2L+1)$, our looser bound is given by

$$\underline{C} \geq \frac{N-P(2L+1)}{N}E\left[\log\left(1+\rho_{\text{eff}}\lambda_{\min}\right)\right] := \underline{C}_a \tag{43}$$

where $\lambda_{\min} = \min\{\lambda_k\}_{k=1}^{\overline{N}_s}$.

C. Optimal Number of Sub-blocks

In Proposition 2, we established that the optimal number of pilots per sub-block is $N_{b,p} = 2L+1$ ($N_{\overline{b},p} = 1$). In this subsection, we will consider what the optimal number of sub-blocks is per transmission block, i.e., how often we should insert the training sub-blocks.

To obtain the optimal number of sub-blocks P in (43), for fixed $N, \mathcal{P}_s$, and $\mathcal{P}_b$, we need to treat P as a continuous variable. Then, we can differentiate $\underline{C}_a$ with respect to P to obtain

$$\begin{aligned}N\frac{\partial\underline{C}_a}{\partial P} &= -(2L+1)E[\log(1+\rho_{\text{eff}}\lambda_{\min})]\\ &\quad + P\overline{N}_sE\left[\log(e)\frac{\rho_{\text{eff}}\lambda_{\min}}{1+\rho_{\text{eff}}\lambda_{\min}}\frac{(2L+1)\sigma_w^2}{\mathcal{P}_s\sigma_{\tilde{h}}^2+P\overline{N}_s\sigma_w^2}\right]\\ &\leq -E\left[\log(e)\frac{\rho_{\text{eff}}\lambda_{\min}(2L+1)}{1+\rho_{\text{eff}}\lambda_{\min}}\right.\\ &\quad \left.-\log(e)\frac{\rho_{\text{eff}}\lambda_{\min}}{1+\rho_{\text{eff}}\lambda_{\min}}\frac{(2L+1)P\overline{N}_s\sigma_w^2}{\mathcal{P}_s\sigma_{\tilde{h}}^2+P\overline{N}_s\sigma_w^2}\right]\\ &= -E\left[\log(e)\frac{\rho_{\text{eff}}\lambda_{\min}}{1+\rho_{\text{eff}}\lambda_{\min}}\frac{(2L+1)\mathcal{P}_s\sigma_{\tilde{h}}^2}{\mathcal{P}_s\sigma_{\tilde{h}}^2+P\overline{N}_s\sigma_w^2}\right] < 0\end{aligned}$$

where in the second step, we used the inequality $\ln(1+x) \geq x/(1+x)\ \forall x > 0$. Since $\partial\underline{C}_a/\partial P < 0$, to achieve the maximum lower bound on the channel capacity, we need to take P as small as possible. Moreover, in order to guarantee the condition in (10) with $N_{b,p} = 2L+1$, we need $P \geq Q+1$. This implies that the optimal number of sub-blocks is $P = Q+1$. Hence, we have established the following proposition.

Proposition 4: Consider transmission of information blocks of length N through a time- and frequency-selective random channel modeled as in (1). If C1)–C4) are satisfied, and a fixed power is allocated to the training symbols, then the lower bound given in (43) is maximized if and only if the number of training sub-blocks is $P = Q + 1$.

Although this result is derived for the looser bound $\underline{C}_a$ in (43), it is also true for the $\underline{C}$ in (40). An intuitive explanation is that as P increases, the performance of channel estimation does not improve, but the number of information symbols decreases, causing $\underline{C}$ to decrease as well. When $P \leq Q + 1$, the mutual information suffers from unreliable channel estimation since the condition in (10) is not satisfied. Note that now, the number of pilot symbols is $(Q+1)(2L+1)$, which is the smallest possible since $P = Q + 1$ [c.f. (10)].

D. Optimal Power Allocation

Up to this point, we have considered that the total power is fixed. Based on this, we have derived that the pilot symbols must be equi-powered and equi-spaced. In this subsection, we will find the optimal allocation of the total power between information symbols and pilots.

Consider the total power $\mathcal{P} = \mathcal{P}_s + \mathcal{P}_b$, and define $\mathcal{P}_s := \alpha\mathcal{P}$; thus, $\mathcal{P}_b = (1-\alpha)\mathcal{P}$ for some $\alpha \in (0, 1)$. From (39), it is easy to verify that $\sigma^2_{\hat{H}} = \overline{N}_s(1-\sigma^2_{\tilde{h}})$, where $\sigma^2_{\tilde{h}}$ is given as [c.f. (30)]

$$\sigma^2_{\tilde{h}} = \sum_{l=0}^{L}\sum_{q=0}^{Q} \frac{\sigma^2_{q,l}\sigma^2_w}{\sigma^2_w + \mathcal{P}_b\sigma^2_{q,l}}. \tag{44}$$

Thus, we can rewrite the effective SNR in (42) as

$$\rho_{\text{eff}} = \frac{\frac{\alpha\mathcal{P}}{Q+1}(1-\sigma^2_{\tilde{h}})}{\sigma^2_w + \frac{\alpha\mathcal{P}}{\overline{N}_s(Q+1)}\sigma^2_{\tilde{h}}}. \tag{45}$$

It is difficult to find an optimal power allocation factor α that does not depend on any CSI directly from (45) because $\sigma^2_{\tilde{h}}$ depends on $\sigma^2_{q,l}$. Therefore, we consider the following three cases:

1) low SNR;
2) high SNR;
3) identical distributed channel taps.

Case i) Low SNR ($\sigma^2_w \gg (1-\alpha)\mathcal{P}\sigma^2_{q,l}$): In this case, we can simplify (44) as $\sigma^2_{\tilde{h}} \gtrapprox 1 - (1-\alpha)\mathcal{P}/\sigma^2_w \sum_{l=0}^{L}\sum_{q=0}^{Q}\sigma^4_{l,q} \gtrapprox 1 - (1-\alpha)\mathcal{P}/\sigma^2_w$. Plugging this result into (45), we obtain

$$\rho_{\text{eff}} \approx \frac{\overline{N}_s\mathcal{P}^2\alpha(1-\alpha)}{\sigma^4_w\overline{N}_s(Q+1) + \alpha\mathcal{P}(\sigma^2_w - (1-\alpha)\mathcal{P})}. \tag{46}$$

The optimal power allocation factor α can be obtained by differentiating ρ_{eff} with respect to the variable α and finding the zero of this differential. Note that α belongs to the range $(0, 1)$. Thus, for this case, we find that

$$\alpha_{low} = 1/2. \tag{47}$$

Case ii) High SNR ($(1-\alpha)\mathcal{P}\sigma^2_{q,l} \gg \sigma^2_w$): In this case, we have from (44) that $\sigma^2_{\tilde{h}} \approx (L+1)(Q+1)\sigma^2_w/\mathcal{P}_b$, and thus, we can rewrite the effective SNR in (42) as

$$\rho_{\text{eff}} = \frac{\frac{\alpha\mathcal{P}}{Q+1}\left(1 - \frac{(L+1)(Q+1)\sigma^2_w}{(1-\alpha)\mathcal{P}}\right)}{\sigma^2_w + \frac{\alpha\mathcal{P}(L+1)\sigma^2_w}{\overline{N}_s(1-\alpha)\mathcal{P}}}.$$

TABLE I
SUMMARY OF DESIGN PARAMETERS

Parameters	Optimal training
Placement of information symbols	Equally long information sub-blocks (length $\tilde{N}_s$)
Placement of training symbols	Equally long training sub-blocks (length $\tilde{N}_b$)
Structure of training sub-blocks	$\mathbf{b}_p = [\mathbf{0}^T_L\ b\ \mathbf{0}^T_L]^T, \forall p$
Number of Training symbols	$2L+1$ per sub-block
Number of Sub-blocks	$Q+1$ training and $Q+1$ information sub-blocks
Power Allocation	$\alpha = 1/(1+\sqrt{(L+1)/\tilde{N}_s})$

After differentiating ρ_{eff} with respect to α, we find that at high SNR, the optimal power allocation factor is

$$\alpha_{high} = \frac{1 - \left(\frac{L+1}{\overline{N}_s} + \frac{(L+1)(Q+1)\sigma^2_w}{\mathcal{P}}\left(1 - \frac{L+1}{\overline{N}_s}\right)\right)^{1/2}}{1 - \frac{L+1}{\overline{N}_s}}. \tag{48}$$

When the SNR $\mathcal{P}/((L+1)(Q+1)\sigma^2_w) \to \infty$, we have

$$\alpha_\infty = \frac{1}{1+\sqrt{(L+1)/\overline{N}_s}} \tag{49}$$

which coincides with the result in [21].

Case iii) Identical Distributed Channel Coefficient ($\sigma^2_{q,l} = 1/((L+1)(Q+1))$): In this case, we can rewrite (44) as $\sigma^2_{\tilde{h}} = (L+1)(Q+1)\sigma^2_w/(\mathcal{P}_b + (Q+1)(L+1)\sigma^2_w)$. Plugging this simplified $\sigma^2_{\tilde{h}}$ into (45), we obtain

$$\rho_{\text{eff}} = \frac{\mathcal{P}^2\overline{N}_s}{\sigma^2_w(Q+1)} \cdot \frac{\alpha(1-\alpha)}{\overline{N}_s((L+1)(Q+1)\sigma^2_w + \mathcal{P}) - \alpha\mathcal{P}(\overline{N}_s - (L+1))}.$$

Similar to the previous two cases, after differentiating ρ_{eff} with respect to α, we obtain that

$$\alpha_{\text{iid}} = \frac{\beta - \left(\beta^2 - \left(1 - \frac{L+1}{\overline{N}_s}\right)\beta\right)^{1/2}}{1 - \frac{L+1}{\overline{N}_s}} \quad \text{with } \beta = 1 + (L+1)(Q+1)\sigma^2_w/\mathcal{P}. \tag{50}$$

When $\mathcal{P}/((L+1)(Q+1)\sigma^2_w) \to \infty$, α_{iid} converges to α_∞ in (49). When $\mathcal{P}/((L+1)(Q+1)\sigma^2_w) \to 0$, $\alpha_{\text{iid}} \to 1/2$.

Proposition 5: Suppose that C1)–C4) hold true and that the SNR is sufficiently high. Under A1)–A4) and for a fixed $\overline{N}_s$, the lower bound on average capacity is maximized with the MMSE channel estimator when the power allocation factor α is given by (47), (48), or (50).

Our optimal PSAM parameters are summarized for convenience in Table I, and the structure of each transmission block $\mathbf{u}$ is depicted in Fig. 4. The optimal pilot insertion strategy is neat in its simplicity and can be equivalently implemented by one interleaver. Fig. 4 depicts this process with a block diagram, where the vertical arrow in each interleaver block denotes a read-out operation, whereas the horizontal arrow indicates a write-in operation. During each transmission burst, we first generate information-bearing blocks of length $\overline{N}_s(Q+1)$ and then feed them

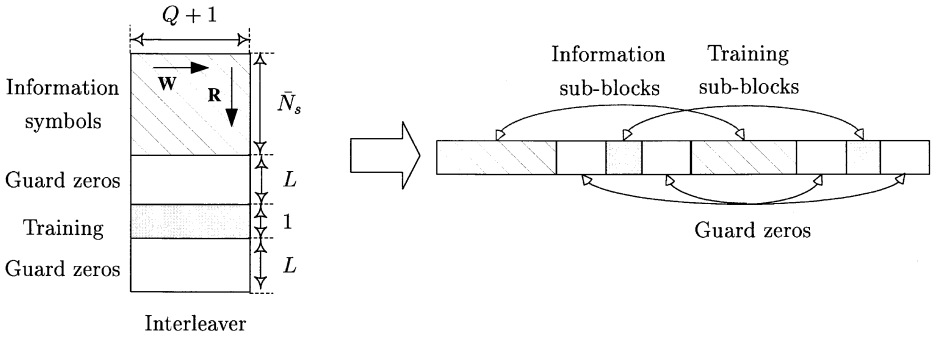


Fig. 4. Transmission design.

to the interleaver (the grey shaded box) as an $\overline{N}_s \times (Q+1)$ matrix, followed by $L \times (Q+1)$ guard zeros (the blank box), $Q+1$ pilot symbols (the grey box), and another $L \times (Q+1)$ guard zeros (the blank box).

In summary, we have designed an optimal training scheme to minimize the channel MMSE and maximize the average capacity. Note that the structure and coding scheme of the information symbols and the optimal block length N have not been touched when the average capacity is considered. On the other hand, the channel parameters Q and L are also related to the diversity order (performance) provided by the channel [3], [18], [24]. Other optimizing criteria such as BER and outage capacity are possible, but their study goes beyond the scope of this paper.

VI. Special Cases and Sampling Interpretation

So far, our entire analysis applies to general doubly selective channels obeying the BEM model. In this section, we will consider two special cases, namely, frequency-selective and time-selective channels. We will first link our results with those in [1], [5], [14], [21], [22], and [30] and show that the latter are subsumed as special cases in our general results here. Later on, we will provide a time-frequency sampling interpretation for our optimal PSAM in (34).

A. Frequency-Selective Channels

Frequency-selective channels exhibit no (or negligible) variation during each transmitted block and correspond to setting $Q = 0$ in (1). Hence, the optimum number of sub-blocks is $Q+1 = 1$, and the transmitted block $\mathbf{u}$ in (34) reduces to

$$\mathbf{u} = [\mathbf{s}^T \ \mathbf{0}_L^T \ b \ \mathbf{0}_L^T]^T \tag{51}$$

where we removed the sub-script p for obvious reasons. Notice that $\mathbf{u}$ in (51) has the same structure as the design in [1, Th. 3], which implies that [1] is subsumed by our design for doubly selective channels. On the other hand, [21] used an affine mapping to represent $\mathbf{s}$ and $\mathbf{b}$. The transmission in (51) can also be written in such an affine form. To show this, let us define matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ as sub-matrices of $\mathbf{I}_{\overline{N}_s+L+1}$, formed by the first $\overline{N}_s$, and the last $L+1$ columns of $\mathbf{I}_{\overline{N}_s+L+1}$, respectively. In addition, let $\mathbf{T}_{zp} := [\mathbf{I}_{\overline{N}_s+L+1}, \mathbf{0}_{(\overline{N}_s+L+1)\times L}]^T$ be the matrix implementing a zero padding operation that pads L zeros when left-multiplying an $(\overline{N}_s+L+1)\times 1$ block. With these notational conventions, it is easy to verify that (51) is equivalent to an affine mapping with transmitted blocks

$$\mathbf{u} = \mathbf{T}_{zp}\mathbf{P}_1\mathbf{s} + \mathbf{T}_{zp}\mathbf{P}_2\mathbf{b}. \tag{52}$$

Our main difference with [1] and [21] is that a cyclic prefix (CP) is employed in [1] and [21] to eliminate IBI while we use zero-padding (ZP). It is interesting that the optimal number of redundant symbols is $2L+1$ for both ZP- and CP-based training designs. Therefore, the bandwidth efficiency

$$\eta = 1 - \frac{2L+1}{N}$$

is the same. Note that in [1] and [21], it is claimed that the optimal number of training symbols is $L+1$, which does not include the cyclic prefix that is needed to avoid IBI. Furthermore, although the power allocation parameter α in (50) and [1] and [21] are identical, they mean different things. Due to the CP, α in [1] and [21] corresponds to the effective information power over the "total" power that excludes the CP. However, in our setup, α corresponds to the ratio of signal power over the total power per block since we use ZP instead of CP to eliminate IBI. So, for a fixed total power per block, our ZP-scheme results in higher effective $\mathcal{P}_s$ and $\mathcal{P}_b$ than the CP-scheme. In this sense, we deduce that the ZP-scheme provides higher average capacity than the CP-scheme does, with the same bandwidth efficiency. As N_s increases, the difference between ZP- and CP-based training decreases. In the simulations section, we will further re-enforce this point.

B. Time-Selective Channels

In time-selective channels, the delay spread can be ignored, and the channel order $L = 0$ must be set in (1). In this case, the transmitted block $\mathbf{u}$ in (34) becomes

$$\mathbf{u} = [\mathbf{s}_1^T \ b \ \mathbf{s}_2^T \ b \ \cdots \mathbf{s}_{Q+1}^T \ b]^T, \qquad b := \sqrt{\overline{\mathcal{P}}_b}. \tag{53}$$

The pilot symbols are inserted equi-spaced and equi-powered. This result coincides with the results in [5] and [22]. Note that in [5], periodic insertion is motivated by uniform sampling arguments. Comparing (53) with [1] and [21], we can observe the duality between periodic insertion of pilots tones in orthogonal

frequency division multiplexing (OFDM) for frequency-selective channels and the PSAM for time-selective channels. There is, however, a notable difference between our scheme in (53) and the optimal design in [22]. In [22], the optimal distance between two consecutive pilots is $\lfloor 1/(2f_{\max}T_s)\rfloor$, where $\lfloor\cdot\rfloor$ denotes integer floor. In contrast, we find the optimal number of pilot symbols per superblock to be $(Q+1)$ since we adopt the BEM as our channel model. A natural question is whether these two designs are related. In the following, we will show that these two designs are in fact equivalent.

Since we rely on the BEM in (1), for a fixed block length N, our Q is defined as

$$Q = 2\lceil f_{\max}T_sN\rceil \gtrapprox 2f_{\max}T_sN. \tag{54}$$

Plugging $N=(Q+1)(\overline{N}_s+1)$ into the inequality (54), we obtain

$$\overline{N}_s+1 \lessapprox \frac{Q}{Q+1}\,\frac{1}{2f_{\max}T_s}.$$

When $Q\gg 1$, i.e., when the block length N is sufficiently large, we find that the distance between two consecutive pilot symbols is $\overline{N}_s+1=\lfloor 1/(2f_{\max}T_s)\rfloor$. Since [22] obtained this optimal distance based on a general time-varying channel model, while we started from the BEM, the equivalence that we just established also corroborates the validity of our BEM.

C. Time-Frequency Sampling Interpretations

For time-selective channels, it is well known that the optimal PSAM samples uniformly the channel in the time domain via periodic insertion of pilot symbols b [5]. Indeed, starting from the scalar input–output relationship for the training samples $y_b(i)=h(i)b+w_b(i)$, one can estimate the channel as $\hat{h}(i)=y_b(i)/b$. In a dual fashion, for frequency-selective channels, optimal PSAM with cyclic prefix samples uniformly the channel in the frequency-domain via periodic insertion of pilot tones $\tilde{b}$ [1], [21]. The input–output relationship now becomes $\tilde{y}_b(q)=\tilde{h}(q)\tilde{b}+\tilde{w}_b(q)$, where $\tilde{y}_b(q)$ denotes the received sample at the qth frequency bin after fast Fourier transform (FFT) processing; similarly, $\tilde{h}(q)$ is the channel transfer function at the qth bin. Channel estimates are now formed in the frequency-domain as $\hat{\tilde{h}}(q)=\tilde{y}_b(q)/\tilde{b}$.

For doubly selective channels, we can view the BEM coefficient $h_q(l)$ in (1) as the two-dimensional (2-D) channel sample at the (qth frequency bin, lth lag or time-slot). We wish to show in this subsection that our optimal PSAM in (34) enables 2-D sampling and estimation of our time-frequency selective channel. Intuitively thinking, the Kronecker deltas in (34) surrounded by zero-guards implement time-domain sampling with pilot symbols; furthermore, the fact that these deltas are periodically inserted implies that they are also equivalent to Kronecker deltas in the frequency-domain and thus serve as pilot tones as well. To solidify this intuition, observe first that with our optimal PSAM in (34), the matrices $\mathbf{B}_p$ in (12) become all equal to $\sqrt{\overline{\mathcal{P}}_b}\mathbf{I}_{L+1}$. Let us now select the $Q+1$ entries from $\mathbf{y}_b$ in (12) with indices $\{(q+1)(l+1)\}_{q=0}^{Q}$ for a fixed lag l. With our optimal $\mathbf{B}_p=\sqrt{\overline{\mathcal{P}}_b}\mathbf{I}_{L+1}$, this allows one to write the input–output training relationship (12) for each $l\in[0,\,L]$ as

$$\mathbf{y}^b(l):=\begin{bmatrix} y_0^b(l)\\ y_1^b(l)\\ \vdots\\ y_Q^b(l)\end{bmatrix}=\sqrt{\overline{\mathcal{P}}_b}\mathbf{F}_{Q+1}^{\mathcal{H}}\mathbf{D}(l)\begin{bmatrix} h_0(l)\\ h_1(l)\\ \vdots\\ h_Q(l)\end{bmatrix}+\mathbf{w}^b(l) \tag{55}$$

where $\mathbf{F}_{Q+1}$ denotes the $(Q+1)$-point FFT matrix with entries $[\exp(-j2\pi(m-1)(n-1)/(Q+1))]_{m,n}$, and $\mathbf{D}(l):=\text{diag}[\exp(j\omega_0(\overline{N}_s+L+l)),\exp(j\omega_1(\overline{N}_s+L+l)),\ldots,\exp(j\omega_Q(\overline{N}_s+L+l))]$. The presence of the inverse FFT matrix in (55) corroborates our intuition that the optimal training in (34) contains pilot tones as well. Let us now concatenate equations like (55) with $l\in[0,\,L]$ to form the $L+1$ columns of the $(Q+1)\times(L+1)$ matrix $\mathbf{Y}_b:=[\mathbf{y}^b(0)\ \ \mathbf{y}^b(1)\cdots\mathbf{y}^b(L)]$. Notice that the matrix $\mathbf{Y}_b$ contains all the training-based received data from $\mathbf{y}_b$ in (12) arranged in a 2-D format. If $\tilde{\mathbf{Y}}_b:=\mathbf{F}_{Q+1}\mathbf{Y}_b/(Q+1)$ denotes the FFT of this 2-D received data array, we can express the training input–output relationship after FFT processing as

$$\tilde{\mathbf{Y}}_b=\sqrt{\overline{\mathcal{P}}_b}\mathbf{D}(0)\begin{bmatrix}1 & e^{j\omega_0} & \cdots & e^{j\omega_0L}\\ 1 & e^{j\omega_1} & \cdots & e^{j\omega_1L}\\ \vdots & \cdots & \cdots & \vdots\\ 1 & e^{j\omega_Q} & \cdots & e^{j\omega_QL}\end{bmatrix}\odot\begin{bmatrix}h_0(0) & \cdots & h_0(L)\\ h_1(0) & \cdots & h_1(L)\\ \vdots & \cdots & \vdots\\ h_Q(0) & \cdots & h_Q(L)\end{bmatrix}+\tilde{\mathbf{W}}_b \tag{56}$$

where $\odot$ denotes the Hadamard product. In scalar form, (56) yields

$$\begin{aligned}\tilde{y}_q^b(l)&=\tilde{b}_q(l)h_q(l)+\tilde{w}_q^b(l),\\ \tilde{b}_q(l)&:=\sqrt{\overline{\mathcal{P}}_b}e^{j\omega_q(\overline{N}_s+L+l)}\end{aligned} \tag{57}$$

which proves that indeed our optimal PSAM samples the BEM in time-frequency to enable estimation of the doubly selective channel via: $\hat{h}_q(l)=\tilde{y}_q^b(l)/\tilde{b}_q(l)$. In fact, our optimal training sequence in (34) is precisely what one needs to obtain the channel model that is assumed *a fortiori* in [14]. Interestingly, starting from the continuous-time channel $h(t;\ \tau)$, and following the steps in [18] to obtain our discrete-time equivalent BEM in (1), one can verify that our time and frequency sampling rates satisfy the 2-D sampling theorem in [14]. Because the latter did not adopt the BEM, this equivalence further confirms the validity of the BEM.

VII. Numerical Examples

We now present test cases to validate our analysis and design. Unless otherwise mentioned, in all test cases, the transmitted

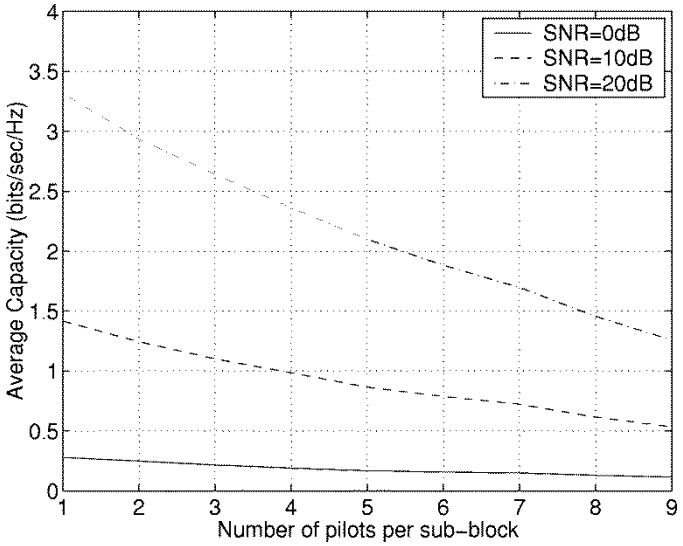


Fig. 5. Capacity versus number of pilots ($\overline{N}_b$).

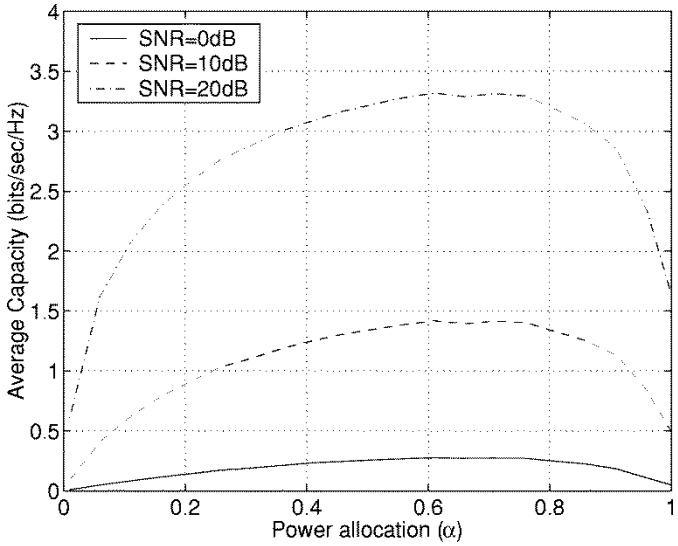


Fig. 6. Power ratio allocation.

block size is $N = 63$, the number of information symbols $N_s = 42$, and the modulation is QPSK. The doubly selective channel model is generated using the following parameters:

- carrier frequency $f_0 = 2$ GHz;
- sampling period $T_s = 53.6$ μs;
- mobile speed $v_{\max} = 160$ km/hr.

Thus, the maximum frequency shift is found to be $f_{\max} \approx 296.30$ Hz. With these parameters, we find that $Q = 2$. Our channel order is $L = 3$. All the channel coefficients $h_q(l)$ are generated as independent, standardized, complex Gaussian random deviates. The multipath intensity profile is selected as $\phi_c(\tau) = \exp(-0.1\tau/T_s)$, $\forall q$, and the Doppler power spectrum is chosen as $S_c(f) = \left(\pi\sqrt{f_{\max}^2 - f^2}\right)^{-1}$ when $f \leq f_{\max}$; otherwise, $S_c(f) = 0$, $\forall l$. We define the variance of $h_q(l)$ as $\sigma_{q,l}^2 := \gamma\phi_c(lT_s)S_c(q/(NT_s))$, where $\gamma := (\sum_{l,q}\phi_c(lT_s)S_c(q/(NT_s)))^{-1}$ denotes the normalizing factor. The signal-to-noise ratio (SNR) is defined as $\mathcal{P}/(N - 2L(Q+1))/\sigma_w^2$.

Test Case 1 (Optimal PSAM Parameters): Two parameters will be tested in this example. The first one is the number of the nonzero pilot symbols $N_{\overline{b},p}$ in C2′). We let $N_{\overline{b},p} = N_{\overline{b}}$, $\forall p$ and adopt all the other parameters in Table I while changing $N_{\overline{b}}$. Fig. 5 depicts the lower bound on the average capacity (40) versus $N_{\overline{b}}$. It can be seen that the capacity bound decreases monotonically as $N_{\overline{b}}$ increases for each SNR value considered (0, 10, and 20 dB). Furthermore, we notice that as the SNR increases, the effect of $N_{\overline{b}}$ increases. The result in Fig. 5 validates the claim in Proposition 3.

Another important parameter we want to test here is the power allocation factor α. We depict the lower bound on the average capacity versus α in Fig. 6. When α is too small (near 0), the average capacity is small since the information symbols do not have enough power to combat AWGN. When α is too large (near 1), the average capacity is also small since the training symbols do not have enough power to provide reliable channel estimation. From (48), the optimal $\alpha \approx 0.65$ in our setup is also verified by inspecting the maximum in Fig. 6.

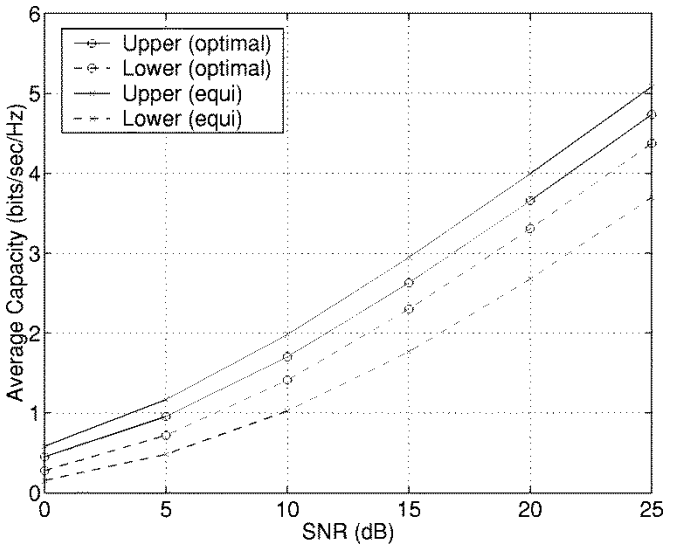


Fig. 7. Optimal power allocation versus equi-powered allocation.

Test Case 2 (Comparison With Equi-Powered PSAM): To emphasize the importance of power allocation, we compare our optimal design (in Table I) with a PSAM design having $\overline{\mathcal{P}}_b = \overline{\mathcal{P}}_s$ and the other parameters selected according to Table I. For this case, the power allocation factor is $\alpha = \overline{N}_s/(1 + \overline{N}_s) \approx 0.93$. From (48), the optimal $\alpha \approx 0.65$. Fig. 7 depicts the lower and upper bounds for both cases. We note that i) for the optimal allocation, the lower bound is closer to the upper bound than for this equi-powered PSAM; therefore, optimal power allocation pays off, and ii) the lower bound for the optimal PSAM is higher than that of equi-powered PSAM since more power is allocated for training in the optimal case. Similar reasoning explains why the upper bound of equi-powered PSAM is higher than that of the optimal PSAM. To further compare the performance of these two cases, we depict the BER versus SNR in Fig. 8. It can be observed that compared with the equi-powered PSAM, the optimal scheme gains 3 dB at 10^{-2}. In Fig. 8, we also plot the

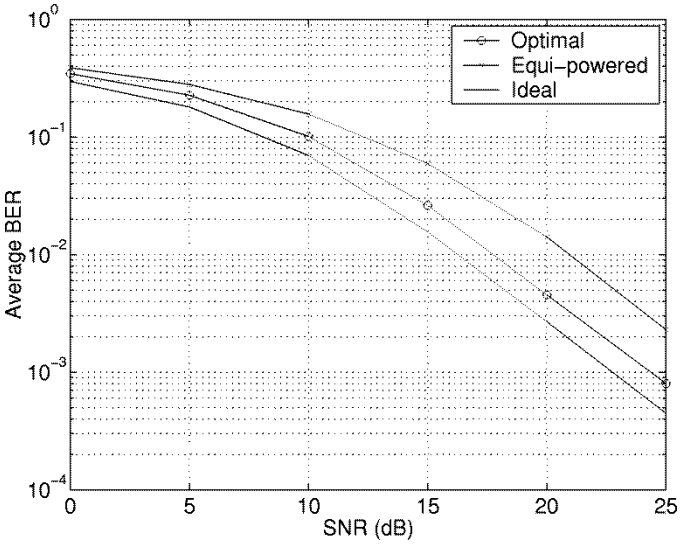


Fig. 8. BER curves for optimal versus equi-powered transmissions.

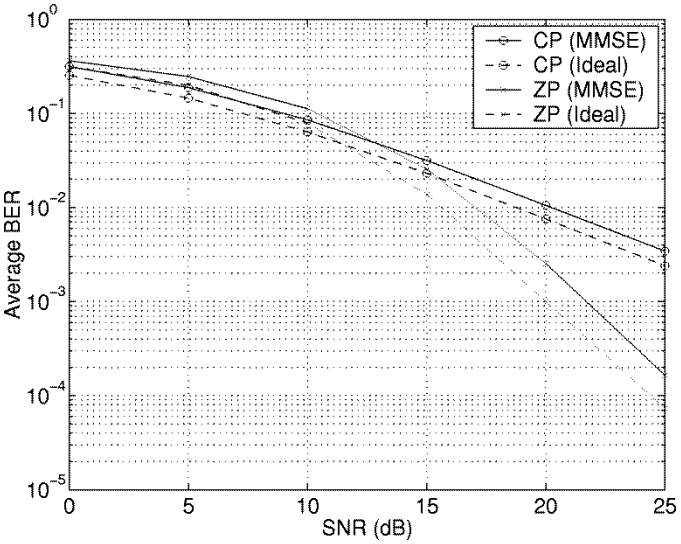


Fig. 10. CP versus ZP for frequency-selective channels.

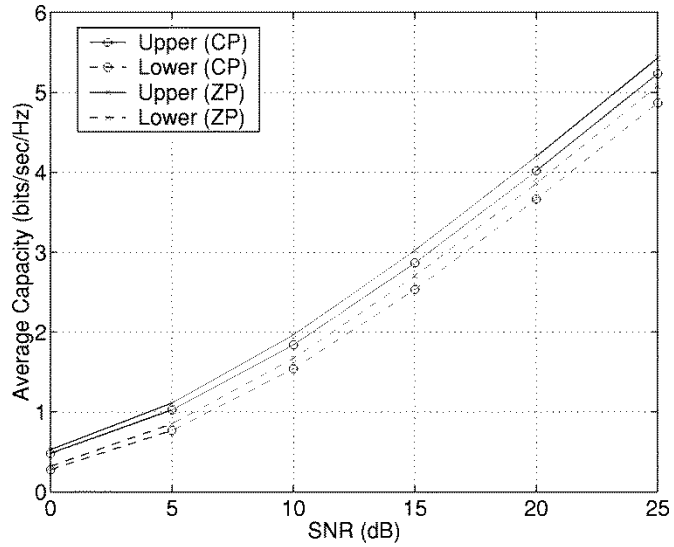


Fig. 9. CP versus ZP for frequency-selective channels.

ideal case with perfect channel estimates. The SNR penalty for channel estimation error is only about 1.5 dB if we adopt the optimal α.

Test Case 3 (Comparison of ZP in (51) With CP in [1] and [21]): This test case is designed to compare our scheme in Section VI-A with [1] and [21]. The channel is frequency-selective with independent and identically distributed (i.i.d.) taps. The channel order $L = 7$, and each tap is a zero mean Gaussian random variable with variance $1/(L+1)$. The number of information symbols per block is $\overline{N}_s = 48$, and the block length $N = \overline{N}_s + 2L + 1$. Therefore, for CP-based training, the CP length is L. The total power per block is fixed to $\mathcal{P}$. Hence, the power ratio allocated between information symbols and training symbols for the CP-based scheme is $\mathcal{P}(\overline{N}_s + L + 1)/N$. Fig. 9 depicts the average capacity bounds for both ZP- and CP-based alternatives. Here, SNR $:= \mathcal{P}/(\overline{N}_s + 1)$. For ZP-based training, the capacity upper and lower bounds are plotted using (21) and (40) with $Q = 0$. For CP-based training, the capacity bounds are plotted according to [21]. Fig. 9 depicts the average capacity bounds for CP- and ZP-based schemes. We notice that the bounds (either upper or lower) for ZP are consistently greater than those of CP, which is partially due to the power loss incurred by the CP.

Although BER is not our design criterion, it is the ultimate performance metric for all communication systems. Therefore, we plot BER versus SNR in Fig. 10. In the same figure, the ideal cases corresponding to perfect channel estimates are also plotted as benchmarks (the dashed lines). We computed MMSE channel estimates based on pilot symbols and used zero-forcing (ZF) equalization for symbol detection in both cases. From Fig. 10, we observe that i) ZP outperforms CP at high SNR, whereas CP has about a 2-dB advantage at BER $= 0.1$; ii) from the slopes of the curves, we notice that CP offers lower diversity order than ZP; and iii) for both cases, the penalty for inaccurate channel state information is about 1.5 dB.

VIII. Concluding Remarks

Optimal PSAM was designed for LMMSE estimation of doubly selective channels by maximizing a lower bound on the average capacity while at the same time minimizing the mean-square channel estimation error. It turned out that the optimal training strategy consists of equi-spaced and equi-powered pilot symbols surrounded by a number of zeros dictated by the channel's delay-spread and inserted periodically with a period dictated by the channel's Doppler-spread. The design enabled a time-frequency sampling of the channel and was shown to subsume time- or frequency-selective channel estimation as special cases.

Our future research will target: i) combining the maximum diversity design in [18] with our optimal training herein to further increase the overall system performance, and ii) extending the optimal training here to space-time coded multiantenna links that encounter doubly selective fading effects.

APPENDIX A
PROOF OF LEMMA 1

Based on the definition of mutual information, we have

$$\mathcal{I}(\mathbf{y}'_s;\ \mathbf{s}|\mathbf{H}) = \mathcal{H}(\mathbf{s}|\mathbf{h}) - \mathcal{H}(\mathbf{s}|\mathbf{h},\ \mathbf{y}'_s)$$

where $\mathcal{H}(\cdot)$ denotes entropy. Since $\mathbf{y}'_s = \mathbf{H}_s\mathbf{s} + \mathbf{w}$, where $\mathbf{w}$ is AWGN, we have that the random variable $\mathbf{s}$ conditioned on $\mathbf{h}$, and $\mathbf{y}'_s$ is Gaussian distributed. Thus, the entropy $\mathcal{H}(\mathbf{s}|\mathbf{h},\ \mathbf{y}'_s)$ can be expressed as

$$\mathcal{H}(\mathbf{s}|\mathbf{h},\ \mathbf{y}'_s) = \log(\det(\pi e\mathbf{R}_{\mathbf{s}|y'_s,\,h}))$$

where $\mathbf{R}_{s|h,\,y'_s} = \left(\mathbf{R}_s^{-1} + (1/\sigma_w^2)\mathbf{H}_s^{\mathcal{H}}\mathbf{H}_s\right)^{-1}$. For a given $\mathbf{R}_s$, we know that $\mathcal{H}(\mathbf{s}|\mathbf{h})$ is maximized when $\mathbf{s}$ is Gaussian [2, p. 143, Th. 3.8], i.e.,

$$\mathcal{H}(\mathbf{s}|\mathbf{h},) \le \log(\det(\pi e\mathbf{R}_s))$$

with equality if and only if the random variable $\mathbf{s}$ is complex Gaussian distributed. Hence, we can upper bound the mutual information $\mathcal{I}(\mathbf{y}'_s;\mathbf{s}|\mathbf{h})$ as

$$\mathcal{I}(\mathbf{y}'_s;\ \mathbf{s}|\mathbf{h}) \le \log\left(\det\left(\mathbf{I} + \frac{1}{\sigma_w^2}\,\mathbf{H}_s\mathbf{R}_s\mathbf{H}_s^{\mathcal{H}}\right)\right)$$

and the average capacity upper-bound in (20) can be written as in (21). ■

APPENDIX B
PROOF OF LEMMA 2

Starting from the conditional mutual information $\mathcal{I}(\mathbf{y}'_s;\ \mathbf{s}|\hat{\mathbf{h}})$, we have

$$\mathcal{I}(\mathbf{y}'_s;\ \mathbf{s}|\hat{\mathbf{h}}) = \mathcal{H}(\mathbf{s}|\hat{\mathbf{h}}) - \mathcal{H}(\mathbf{s}|\hat{\mathbf{h}},\ \mathbf{y}'_s). \tag{58}$$

Recalling (12), we notice that $\hat{\mathbf{h}}$ does not depend on $\mathbf{s}$. Hence, from A4), we have

$$\mathcal{H}(\mathbf{s}|\hat{\mathbf{h}}) = \log(\det(\pi e\mathbf{R}_s)).$$

Similar to Appendix A, based on [2, p. 143, Th. 3.8], we obtain that

$$\mathcal{H}(\mathbf{s}|\hat{\mathbf{h}},\ \mathbf{y}'_s) \le \log\left(\det\left(\pi e\mathbf{R}_{s|\hat{h},\,y'_s}\right)\right) \tag{59}$$

where the equality holds if and only if $\mathbf{s}$ conditioned on $\hat{\mathbf{h}}$ and $\mathbf{y}'_s$ is Gaussian distributed with covariance matrix $\mathbf{R}_{s|\hat{h},\,y'_s} := E[(\mathbf{s} - E[\mathbf{s}|\hat{\mathbf{h}},\ \mathbf{y}'_s])(\mathbf{s} - E[\mathbf{s}|\hat{\mathbf{h}},\ \mathbf{y}'_s])^{\mathcal{H}}]$. Note that when $\mathbf{v}$ is Gaussian, $\mathbf{s}$ for given $\hat{\mathbf{h}}$, and $\mathbf{y}'_s$ is also Gaussian. Thus, the entropy $\mathcal{H}(\mathbf{s}|\hat{\mathbf{h}},\ \mathbf{y}'_s)$ is maximized [c.f. (59)]. Here, $\mathbf{R}_{s|\hat{h},\,y'_s}$ is the covariance matrix of the MMSE estimator of $\mathbf{s}$ for each realization of $\hat{\mathbf{h}}$. Recall that our model is $\mathbf{y}'_s = \hat{\mathbf{H}}_s\mathbf{s} + \mathbf{v}$. Since $\hat{\mathbf{h}}$ is Gaussian, if the noise $\mathbf{v}$ is also Gaussian, then the LMMSE estimator of $\mathbf{s}$ is an MMSE estimator. Therefore, we can obtain $\mathbf{R}_{s|\hat{h},\,y'_s}$ by the LMMSE estimator of $\mathbf{s}$ as

$$\mathbf{R}_{s|\hat{h},\,y'_s} = \mathbf{R}_s - \mathbf{R}_{sy}\mathbf{R}_y^{-1}\mathbf{R}_{ys} \tag{60}$$

where

$$\begin{aligned}\mathbf{R}_{sy} &:= E[\mathbf{s}(\mathbf{y}'_s)^{\mathcal{H}}|\hat{\mathbf{h}}] = \mathbf{R}_s\hat{\mathbf{H}}_s^{\mathcal{H}} + E[\mathbf{s}\mathbf{v}^{\mathcal{H}}|\hat{\mathbf{h}}]\\ \mathbf{R}_y &:= E[\mathbf{y}'_s(\mathbf{y}'_s)^{\mathcal{H}}|\hat{\mathbf{h}}]\\ &= \hat{\mathbf{H}}_s\mathbf{R}_s\hat{\mathbf{H}}_s^{\mathcal{H}} + \mathbf{R}_v + \hat{\mathbf{H}}_s E[\mathbf{s}\mathbf{v}^{\mathcal{H}}|\hat{\mathbf{h}}] + E[\mathbf{v}\mathbf{s}^{\mathcal{H}}|\hat{\mathbf{h}}]\hat{\mathbf{H}}_s^{\mathcal{H}}.\end{aligned}$$

Because $\mathbf{h}$ is Gaussian and the noise $\mathbf{w}_b$ in (12) is Gaussian, for the LMMSE channel estimator, we have $E[\tilde{\mathbf{h}}|\hat{\mathbf{h}}] = E[\mathbf{h} - \hat{\mathbf{h}}|\hat{\mathbf{h}}] = \mathbf{0}$. Thus, we can verify that

$$E[\mathbf{s}\mathbf{v}^{\mathcal{H}}|\hat{\mathbf{h}}] = E[\mathbf{s}\mathbf{s}^{\mathcal{H}}\tilde{\mathbf{H}}_s^{\mathcal{H}}|\hat{\mathbf{h}}] + E[\mathbf{s}\overline{\mathbf{b}}^{\mathcal{H}}\tilde{\mathbf{H}}_b^{\mathcal{H}}|\hat{\mathbf{h}}] = \mathbf{0}. \tag{61}$$

Taking (61) into account, we can rewrite (60) as

$$\begin{aligned}\mathbf{R}_{s|\hat{h},\,y'_s} &= \mathbf{R}_s - \mathbf{R}_s\hat{\mathbf{H}}_s^{\mathcal{H}}(\hat{\mathbf{H}}_s\mathbf{R}_s\hat{\mathbf{H}}_s^{\mathcal{H}} + \mathbf{R}_v)^{-1}\hat{\mathbf{H}}_s\mathbf{R}_s\\ &= (\mathbf{R}_s^{-1} + \hat{\mathbf{H}}_s^{\mathcal{H}}\mathbf{R}_v^{-1}\hat{\mathbf{H}}_s)^{-1}.\end{aligned}$$

Plugging $\mathbf{R}_{s|\hat{h},\,y'_s}$ into (58), we obtain

$$\mathcal{I}(\mathbf{y}'_s;\ \mathbf{s}|\hat{\mathbf{h}}) \ge \log(\det(\mathbf{I} + \mathbf{R}_s\hat{\mathbf{H}}_s^{\mathcal{H}}\mathbf{R}_v^{-1}\hat{\mathbf{H}}_s)).$$

Therefore, the average capacity in (19) has a lower bound given in (24). ■

APPENDIX C
PROOF OF LEMMA 3

Suppose there are two training schemes leading to correlation matrices $\mathbf{R}_v$ and $\mathbf{R}_{v'}$. Let $\mathbf{\Delta} := \mathbf{R}_v - \mathbf{R}_{v'} \ge \mathbf{0}$ in the positive semi-definite sense, and note that both $\mathbf{R}_v$ and $\mathbf{R}_{v'}$ are full-rank matrices [c.f. (23)]. By eigen-decomposing $\mathbf{\Delta}$, we obtain that $\mathbf{\Delta} = \mathbf{U}_{\mathbf{\Delta}}\mathbf{\Lambda}_{\mathbf{\Delta}}\mathbf{U}_{\mathbf{\Delta}}^{\mathcal{H}}$, where $\mathbf{\Lambda}_{\mathbf{\Delta}}$ includes all nonzero eigenvalues of $\mathbf{\Delta}$, and the columns of $\mathbf{U}_{\mathbf{\Delta}}$ are the corresponding eigenvectors. From the matrix inversion lemma, we have that

$$\begin{aligned}\mathbf{R}_v^{-1} &= \mathbf{R}_{v'}^{-1} - \mathbf{R}_{v'}^{-1}\mathbf{U}_{\mathbf{\Delta}}(\Lambda_{\mathbf{\Delta}}^{-1} + \mathbf{U}_{\mathbf{\Delta}}^{\mathcal{H}}\mathbf{R}_{v'}^{-1}\mathbf{U}_{\mathbf{\Delta}})^{-1}\mathbf{U}_{\mathbf{\Delta}}\mathbf{R}_{v'}^{-1}\\ &:= \mathbf{R}_{v'}^{-1} - \mathbf{\Delta}'\end{aligned} \tag{62}$$

where $\mathbf{\Delta}' \ge \mathbf{0}$, and hence, $\mathbf{R}_v^{-1} \le \mathbf{R}_{v'}^{-1}$. Using the monotonicity of log, and the property $\det(\mathbf{A} + \mathbf{B}) \ge \det(\mathbf{A})$, for $\mathbf{A},\ \mathbf{B} \ge \mathbf{0}$, we infer that at high SNR ($\hat{\mathbf{H}}_s \approx \mathbf{H}_s$), it holds that [c.f. (62)]

$$\begin{aligned}&\log\det\left(\mathbf{I}_{N_s+LP} + \overline{\mathcal{P}}_s\mathbf{R}_v^{-1}\hat{\mathbf{H}}_s\hat{\mathbf{H}}_s^{\mathcal{H}}\right)\\ &\qquad\le \log\det\left(\mathbf{I}_{N_s+LP} + \overline{\mathcal{P}}_s\mathbf{R}_{v'}^{-1}\hat{\mathbf{H}}_s\hat{\mathbf{H}}_s^{\mathcal{H}}\right).\end{aligned} \tag{63}$$

Therefore, minimizing $\mathbf{R}_v$ is equivalent to maximizing $\underline{C}$. ■

APPENDIX D
PROOF OF LEMMA 4

Suppose there are two training schemes with identical $E[\tilde{\mathbf{h}}\tilde{\mathbf{h}}^{\mathcal{H}}]$, leading to correlation matrices $\mathbf{R}_v$ and $\mathbf{R}_{v'}$, as in (23). We observe that the first term on the right-hand side of (23) is identical for the two schemes. If the scheme with $\mathbf{R}_v$ has the first L and the last L entries of $\mathbf{b}_p$ equal to zero (a two-sided zero-guard condition), then the second term in (23) is zero because $\tilde{\mathbf{H}}_b\overline{\mathbf{b}} = \mathbf{0}$ [c.f. (3)]; if the one with $\mathbf{R}_{v'}$ does not have this zero-guard, then it cannot null this positive semi-definite term. Therefore, we obtain $\mathbf{R}_v \leq \mathbf{R}_{v'}$. If $\exists\, N_{b,p} \leq 2L$, then the two-sided zero-guard condition requires that $\mathbf{b}_p = \mathbf{0}$. In this case, $\mathbf{b}_p$ cannot be used for channel estimation; indeed, if $\mathbf{b}_p \neq \mathbf{0}$, then the minimum $\mathbf{R}_v$ cannot be achieved. Hence, we require that $N_{b,p} \geq 2L+1$. ■

APPENDIX E
PROOF OF LEMMA 5

From the proof of Proposition 1, we know that for $N_{s,p} \gg 2L$, we can approximate $\mathbf{R}_v$, as in (29). Then, the lower bound on the average capacity becomes

$$\underline{C} \approx \frac{1}{N}\sum_{p=1}^{P} E\Big[\log\det\big(\mathbf{I}_{N_{s,p}+L} + \overline{\mathcal{P}}_s(\sigma_{\tilde{h}}^2\overline{\mathcal{P}}_s + \sigma_w^2)^{-1} \cdot \hat{\mathbf{H}}_p^s(\hat{\mathbf{H}}_p^s)^{\mathcal{H}}\big)\Big] \text{ bits/s/Hz.}$$

With the definition

$$\underline{C}_p = E\Big[\log\det\big(\mathbf{I}_{N_{s,p}+L} + \overline{\mathcal{P}}_s(\sigma_{\tilde{h}}^2\overline{\mathcal{P}}_s + \sigma_w^2)^{-1} \cdot \hat{\mathbf{H}}_p^s(\hat{\mathbf{H}}_p^s)^{\mathcal{H}}\big)\Big] \text{ bits/s/Hz}$$

it has been shown in [1] that $\underline{C}_p$ is a concave function of $N_{s,p}$. Therefore, we have that

$$\underline{C} \leq \frac{1}{N}\sum_{p=1}^{P} E\Big[\log\det\Big(\mathbf{I}_{\overline{N}_s+L} + \overline{\mathcal{P}}_s(\sigma_{\tilde{h}}^2\overline{\mathcal{P}}_s + \sigma_w^2)^{-1} \cdot \hat{\mathbf{H}}_p^s(\hat{\mathbf{H}}_p^s)^{\mathcal{H}}\Big)\Big] \text{ bits/s/Hz} \quad (64)$$

where $\overline{N}_s := N_s/P$. Equation (64) shows clearly that $\underline{C}$ is maximized when equally long information sub-blocks are designed to have length $N_{s,p} = \overline{N}_s = N_s/P$. ■

APPENDIX F
PROOF OF LEMMA 6

$\forall\, p_1, p_2 \in [1, P]$, and $p_1 \neq p_2$, we can write that

$$\hat{\overline{\mathbf{H}}}_{p_1}^s = [\mathbf{D}_{0,p_1}, \ldots, \mathbf{D}_{Q,p_1}]\begin{bmatrix}\hat{\overline{\mathbf{H}}}_{0,p_1}^s \\ \vdots \\ \hat{\overline{\mathbf{H}}}_{Q,p_1}^s\end{bmatrix}$$

which allows us to factor $\hat{\overline{\mathbf{H}}}_{p_2}$ as

$$\hat{\overline{\mathbf{H}}}_{p_2}^s = [\mathbf{D}_{0,p_1}, \ldots, \mathbf{D}_{Q,p_1}]\begin{bmatrix}e^{j\omega_0(p_2-p_1)(\overline{N}_s+2L+1)}\hat{\overline{\mathbf{H}}}_{0,p_2}^s \\ \vdots \\ e^{j\omega_Q(p_2-p_1)(\overline{N}_s+2L+1)}\hat{\overline{\mathbf{H}}}_{Q,p_2}^s\end{bmatrix}.$$

It can be verified that $\hat{\overline{\mathbf{H}}}_{q,p_1}^s = \hat{\overline{\mathbf{H}}}_{q,p_2}^s, \forall\, q \in [0, Q]$. Since $\mathbf{H}$ is Gaussian [c.f. A2)], the estimate $\hat{\mathbf{h}}$ is also Gaussian, and thus, $\hat{\overline{\mathbf{H}}}_{q,p_1}$ is Gaussian. Because

$$E[\hat{\overline{\mathbf{H}}}_{q,p_1}^s] = E[e^{j\omega_q(p_2-p_1)(\overline{N}_s+2L+1)}\overline{\mathbf{H}}_{q,p_1}^s] = \mathbf{0}$$
$$E[\hat{\overline{\mathbf{H}}}_{q,p_1}^s(\hat{\overline{\mathbf{H}}}_{q,p_1}^s)^{\mathcal{H}}] = E[\overline{\mathbf{H}}_{q,p_1}^s(\overline{\mathbf{H}}_{q,p_1}^s)^{\mathcal{H}}].$$

Because of the Gaussianity, we obtain that $\hat{\overline{\mathbf{H}}}_{q,p_1}^s$ and $e^{j\omega_q(p_2-p_1)(\overline{N}_s+2L+1)}\hat{\overline{\mathbf{H}}}_{q,p_2}^s$ have the same distribution. So do $\hat{\overline{\mathbf{H}}}_{p_1}^s$ and $\hat{\overline{\mathbf{H}}}_{p_2}^s$. Since N is fixed, as $N_{b,p}$ increases, we infer that $\overline{N}_s$ decreases, and so does $\underline{C}$. ■

ACKNOWLEDGMENT

The authors would like to thank Prof. L. Tong of Cornell University for his suggestions on the proof of Lemma 2.

REFERENCES

[1] S. Adireddy, L. Tong, and H. Viswanathan, "Optimal placement of training for unknown channels," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2338–2353, Aug. 2002.
[2] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*. New York: Kluwer/Plenum, 1999.
[3] S. Bhashyam, A. M. Sayeed, and B. Aazhang, "Time-selective signaling and reception for communication over multipath fading channels," *IEEE Trans. Commun.*, vol. 48, pp. 83–94, Jan. 2000.
[4] D. K. Borah and B. T. Hart, "Frequency-selective fading channel estimation with a polynomial time-varying channel model," *IEEE Trans. Commun.*, vol. 47, pp. 862–873, June 1999.
[5] J. K. Cavers, "An analysis of pilot symbol assisted modulation for Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 40, pp. 686–693, Nov. 1991.
[6] ——, "Pilot symbol assisted modulation and differential detection in fading and delay spread," *IEEE Trans. Commun.*, vol. 43, pp. 2206–2212, July 1995.
[7] A. P. Clark and S. Hariharan, "Adaptive channel estimator for an HF radio link," *IEEE Trans. Commun.*, vol. 37, pp. 918–926, Sept. 1989.
[8] L. Davis, I. Collings, and P. Höeher. (2001, May) Joint MAP equalization and channel estimation for frequency-selective and frequency-flat fast-fading channels. *IEEE Trans. Commun.* [Online]. Available: http://www-lns.techfak.uni-kiel.de/ict/.
[9] M. Dong and L. Tong, "Optimal design and placement of pilot symbols for channel estimation," *IEEE Trans. Signal Processing*, vol. 50, pp. 3055–3069, Dec. 2002.
[10] A. Duel-Hallen, S. Hu, and H. Hallen, "Long-range prediction of fading channels," *IEEE Signal Processing Mag.*, pp. 62–75, May 2000.
[11] S. A. Fechtel and H. Meyr, "Optimal parametric feedforward estimation of frequency-selective fading radio channels," *IEEE Trans. Commun.*, vol. 42, pp. 1639–1650, Feb./Mar./Apr. 1994.
[12] G. B. Giannakis and C. Tepedelenlioğlu, "Basis expansion models and diversity techniques for blind identification and equalization of time-varying channels," *Proc. IEEE*, vol. 86, pp. 1969–1986, Nov. 1998.
[13] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, Nov. 1972.
[14] P. Höeher, S. Kaiser, and P. Robertson, "Two-dimensional pilot-symbol-aided channel estimation by Wiener filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Münich, Germany, 1997, pp. 1845–1848.
[15] T. Kailath, "Measurements on time-variant communication channels," *IEEE Trans. Inform. Theory*, vol. IT-8, pp. S229–S236, Sept. 1962.
[16] W. Kozek, "On the transfer function calculus for underspread LTV channels," *IEEE Trans. Signal Processing*, vol. 45, pp. 219–223, Jan. 1997.
[17] W. C. Jakes, *Microwave Mobile Communications*. New York: Wiley, 1974.
[18] X. Ma and G. B. Giannakis, "Maximum-diversity transmissions over doubly selective wireless channels," *IEEE Trans. Inform. Theory*, 2003, to be published. [Online] Available: http://spincom.ece.umn.edu/xiaoli/double-diversity.pdf.
[19] M. Martone, "Wavelet-based separating kernels for sequence estimation with unknown rapidly time-varying channels," *IEEE Commun. Lett.*, vol. 3, pp. 78–80, Mar. 1999.

[20] S. Ohno and G. B. Giannakis, "Optimal training and redundant precoding for block transmissions with application to wireless OFDM," *IEEE Trans. Commun.*, vol. 50, pp. 2113–2123, Dec. 2002.
[21] ——, "Capacity maximizing pilots for wireless OFDM over rapidly fading channels," *IEEE Trans. Inform. Theory*. see also *Proc. Int. Symp. Signals, Syst., Electron.*, pp. 246–249, Tokyo, Japan, July 24–27, 2001, to be published.
[22] ——, "Average-rate optimal PSAM transmissions over time-selective fading channels," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 712–720, Oct. 2002.
[23] J. G. Proakis, *Digital Communications*, Fourth ed. New York: McGraw-Hill, 2001.
[24] A. M. Sayeed and B. Aazhang, "Joint multipath-doppler diversity in mobile wireless communications," *IEEE Trans. Commun.*, vol. 47, pp. 123–132, Jan. 1999.
[25] C. Tepedelenlioğlu and G. B. Giannakis, "Transmitter redundancy for blind estimation and equalization of time- and frequency-selective channels," *IEEE Trans. Signal Processing*, vol. 48, pp. 2029–2043, July 2000.
[26] M. K. Tsatsanis and G. B. Giannakis, "Equalization of rapidly fading channels: Self-recovering methods," *IEEE Trans. Commun.*, vol. 44, pp. 619–630, May 1996.
[27] ——, "Modeling and equalization of rapidly fading channels," *Int. J. Adapt. Contr. Signal Process.*, vol. 10, pp. 159–176, May 1996.
[28] M. K. Tsatsanis and Z. Xu, "Pilot symbol assisted modulation in frequency selective fading wireless channels," *IEEE Trans. Signal Processing*, vol. 48, pp. 2353–2365, Aug. 2000.
[29] J. K. Tugnait and B. Huang, "Second-order statistics-based blind equalization of IIR single-input multiple-output channels with common zeros," *IEEE Trans. Signal Processing*, vol. 47, pp. 147–157, July 1999.
[30] H. Vikalo, B. Hassibi, B. Hochwald, and T. Kailath, "Optimal training for frequency-selective fading channels," in *Proc. Int. Conf. Acoust, Speech, Signal Process.*, vol. 4, Salt Lake City, UT, May 7–11, 2001, pp. 2105–2108.
[31] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: Where Fourier meets Shannon," *IEEE Signal Processing Mag.*, vol. 17, pp. 29–48, May 2000.
[32] Y. Zhang, M. P. Fitz, and S. B. Gelfand, "A performance analysis and design of equalization with pilot aided channel estimation," in *Proc. 47th Veh. Technol. Conf.*, vol. 2, Phoenix, AZ, May 4–7, 1997, pp. 720–724.
[33] ——, "Soft output demodulation on frequency-selective Rayleigh fading channels using AR channel models," in *Proc. Global Commun. Conf.*, vol. 1, Phoenix, AZ, Nov. 3–8, 1997, pp. 327–331.

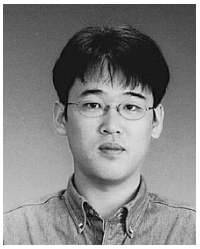
Xiaoli Ma received the B.S. degree in automatic control from Tsinghua University, Beijing, China, in 1998 and the M.S. degree in electrical engineering from University of Virginia, Charlottesville, in 1999. She is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis.

Her research interests include transmitter and receiver diversity techniques for fading channels, communications over time- and frequency-selective channels, complex-field and space-time coding, channel estimation, equalization, and synchronization.

Georgios B. Giannakis (F'97) received the Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1981. From September 1982 to July 1986, he was with the University of Southern California (USC), Los Angeles, where he received the MSc. degree in electrical engineering in 1983, the MSc. degree in mathematics in 1986, and the Ph.D. degree in electrical engineering, also in 1986.

After lecturing for one year at USC, he joined the University of Virginia, Charlottesville, in 1987, where he became a Professor of electrical engineering in 1997. Since 1999, he has been a Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, where he now holds an ADC Chair in Wireless Telecommunications. His general interests span the areas of communications and signal processing, estimation and detection theory, time-series analysis, and system identification—subjects on which he has published more than 150 journal papers, 300 conference papers, and two edited books. His current research topics focus on transmitter and receiver diversity techniques for single- and multiuser fading communication channels, complex-field and space-time coding for block transmissions, multicarrier, and ultrawide band wireless communication systems. He is a frequent consultant for the telecommunications industry.

Dr. Giannakis is the (co-) recipient of four best paper awards from the IEEE Signal Processing (SP) Society (in 1992, 1998, 2000, and 2001). He also received the Society's Technical Achievement Award in 2000. He co-organized three IEEE-SP Workshops, and guest (co-) edited four special issues. He has served as Editor-in-Chief of the IEEE SIGNAL PROCESSING LETTERS, as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE SIGNAL PROCESSING LETTERS, as Secretary of the SP conference board, as member of the SP Publications Board, as member and vice-chair of the Statistical Signal and Array Processing Technical Committee, and as chair of the SP for Communications Technical Committee. He is a member of the Editorial Board for the PROCEEDINGS OF THE IEEE and the steering committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He is a member of the IEEE Fellows Election Committee and the IEEE-SP Society's Board of Governors.



Shuichi Ohno (M'95) received the B.E., M.E., and Dr. Eng. degrees in applied mathematics and physics from Kyoto University, Kyoto, Japan, in 1990, 1992, and 1995, respectively.

From 1995 to 1999, he was a research associate with the Department of Mathematics and Computer Science, Shimane University, Shimane, Japan, where he became an Assistant Professor. He spent 14 months in 2000 and 2001 at the University of Minnesota, Minneapolis, as a visiting researcher. Since 2002, he has been an Associate Professor with the Department of Artificial Complex Systems Engineering, Hiroshima University, Hiroshima, Japan. His current interests are in the areas of signal processing in communication, wireless communications, adaptive signal processing, and multirate signal processing.

Dr. Ohno is a member IEICE. He has been serving as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS since 2001.