Capacity Maximizing MMSE-Optimal Pilots for Wireless OFDM Over Frequency-Selective Block Rayleigh-Fading Channels

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

***Abstract*—The location, number, and power of pilot symbols embedded in multicarrier block transmissions over rapidly fading channels, are important design parameters affecting not only channel estimation performance, but also channel capacity. Considering orthogonal frequency-division multiplexing (OFDM) systems with decoupled information-bearing symbols from pilot symbols transmitted over wireless frequency-selective Rayleigh-fading channels, we show that equispaced and equipowered pilot symbols are optimal in terms of minimizing the mean-square channel estimation error. We also design the number of pilots, and the power distributed between information bearing and pilot symbols, using as criterion a lower bound on the average capacity. Numerical results corroborate our theoretical findings.**

***Index Terms*—Block transmissions, capacity, channel estimation, orthogonal frequency-division multiplexing (OFDM), pilot tones.**

I. Introduction

Block transmissions relying on linear redundant precoding with cyclic prefixed or zero padded blocks have gained increasing interest recently for mitigating frequency-selective multipath effects (see, e.g., [4], [15], [25], [27], and references therein). Redundancy removes interblock interference (IBI) and facilitates (even blind) acquisition of channel state information (CSI) at the receiver. When CSI is available at the transmitter (e.g., through a feedback channel) precoders and decoders can be optimized under various criteria [25]. For example, maximizing mutual information leads to a waterfilling-type transmission that invokes power and bit loading to approach the channel capacity and control the average bit-error rate (BER) [24]. However, rapid variations of the wireless channel render CSI feedback to the transmitter outdated. On the other hand, when CSI is used at the receiver for coherent detection, training sequences are needed to acquire it.

Instead of long training sequences at the beginning of the transmitted record, inserting training symbols during the transmission is known as pilot symbol-aided modulation (PSAM). Since pilots are embedded in each transmitted block, PSAM is suitable for rapidly fading time-selective channels. Known pilot symbols are used not only for channel estimation [6], [11], [12], [19], but also for timing- and frequency-offset synchronization [15], [23]. PSAM is also useful for decision-feedback (DF) equalization of block transmissions [13], with the optimum placement of pilots for DF equalization studied in [1]. Certainly, pilots reduce the information rate, and thus offset utilization of the channel's capacity.

Manuscript received April 30, 2001; revised March 2, 2004. This work was supported by the National Science Foundation under Grant CCR-10105612. The material in this correspondence was presented in part at the International Symposium on Signals, Systems and Electronics (ISSSF'01), Tokyo, Japan, July 2001.

S. Ohno is with the Department of Artificial Complex Systems Engineering, Hiroshima University, Higashi-Hiroshima, 739-8527, Japan (e-mail: o.shuichi@ieee.org).

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

Communicated by P. Narayan, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2004.833365

The effect of imperfect CSI on channel capacity has been investigated for time-selective point-to-point channels [5], [16], [18], and for multiple-antenna communication channels [10], [21], [22], [29]. Noncoherent channel capacity of multiple-antenna communication channels with infinite transmit power was also studied in [30]. For a given channel estimation accuracy, lower and upper bounds on channel capacity are available [10], [16], [18], [21]. Based on a lower bound on channel capacity, optimal training sequences for multiple-antenna wireless communications over frequency-flat block-fading channels were pursued in [10], while the generalized mutual information (GMI) was utilized in [29] to design training sequences, where the optimality of minimum mean-square error (MMSE) channel estimators in terms of GMI was also proven. On the other hand, the accuracy of frequency-selective channel estimators depends on the number, the location, and the power of training symbols inserted per transmitted block [2], [3], [20]. Relying on a capacity bound, the design of optimal training sequences for transmissions over finite impulse response (FIR) frequency-selective channels was dealt with in [2], [3], where it is assumed that the channel taps are independent and identically distributed (i.i.d.), and that the training symbols are taken *a fortiori* to be equipowered.

This correspondence deals with linearly precoded block transmissions with each block containing superimposed training symbols for random FIR channel estimation. We design the number, the location, and the power of pilots. We first develop pilots to minimize the channel estimation error and then maximize the lower bound on the capacity of block Rayleigh multipath channels. After developing our block modeling framework, we review briefly the design constraints for IBI cancellation and low-complexity block-by-block reception in Section II, to enable linear channel estimation that is decoupled from symbol recovery [20]. Relying on the optimality of MMSE estimators [29], we derive pilots that minimize the channel's mean-square error (MSE) in Section III. Subsequently, to maximize the lower bound on channel capacity, we select the number of pilot tones, and determine the power distribution between pilot and information-bearing symbols in Section IV. Simulations are presented in Section V, before concluding the correspondence.

II. Background and Preliminaries

We consider the following discrete-time baseband equivalent model for block transmissions: the information-bearing sequence $s(n)$ is parsed into blocks of size M. Each $M \times 1$ block $\boldsymbol{s}(i)$ is precoded by a tall $\bar{N} \times M$ precoding matrix $\bar{\boldsymbol{A}}$ with complex-valued entries. Selecting $\bar{N} > M$ introduces redundancy, which is known to be beneficial for mitigating the effects of frequency-selective propagation [27], [28]. An $\bar{N} \times 1$ block of training symbols $\bar{\boldsymbol{b}}$, which are also known to the receiver, is added to the precoded information block to obtain

$$\bar{\boldsymbol{u}}(i) = \bar{\boldsymbol{A}}\boldsymbol{s}(i) + \bar{\boldsymbol{b}}. \tag{1}$$

The channel is considered to be frequency-selective linear time-invariant over each received block, but is allowed to vary from block to block. We will only consider block-by-block receiver processing; and we will henceforth omit time dependence of the channel and express the FIR of the discrete-time baseband equivalent channel as $\{h(n)\}$.

The order of the channel is assumed to be upper-bounded by $L(< M)$. At the receiver, we collect $\bar{N}$ noisy samples in an $\bar{N} \times 1$ received vector $\bar{\boldsymbol{x}}(i)$ that can be expressed as (see, e.g., [27])

$$\bar{\boldsymbol{x}}(i) = \boldsymbol{H}_0\bar{\boldsymbol{u}}(i) + \boldsymbol{H}_1\bar{\boldsymbol{u}}(i-1) + \boldsymbol{\eta}(i) \tag{2}$$

where $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ are square Toeplitz channel convolution matrices with first column $[h(0), h(1), \ldots, h(L), 0, \ldots, 0]^T$ and with first row $[0, \ldots, 0, h(L), h(L-1), \ldots, h(1)]$, respectively; and $\boldsymbol{\eta}(i)$ is a zero-mean additive Gaussian noise (AGN) vector. Presence of two successive transmitted blocks in each received block $\bar{\boldsymbol{x}}(i)$ arises due to the channel that has length $\bar{L} := L + 1$. The second term in (2) is referred to as IBI.

To enable low-complexity *block-by-block processing* at the receiver, we will eliminate IBI by inserting at the transmitter and removing at the receiver the so-called cyclic prefix (CP), which is employed also by orthogonal frequency division multiplexing (OFDM)—the basic multicarrier modulation that has been adopted by many standards; see e.g., [8]. To enable insertion and removal of the CP, and thus cancellation of IBI, we select our general design in (1) to satisfy the following.

C1. Matrix $\bar{\boldsymbol{A}}$ and vector $\bar{\boldsymbol{b}}$ are chosen to incorporate the CP; i.e., $\bar{\boldsymbol{A}} = \boldsymbol{T}_{\text{cp}}\boldsymbol{A}$, $\bar{\boldsymbol{b}} = \boldsymbol{T}_{\text{cp}}\boldsymbol{b}$, where $\boldsymbol{A}$ is an $N(:= \bar{N} - L) \times M$ matrix, $\boldsymbol{T}_{\text{cp}} := [[\boldsymbol{0}_{L\times(N-L)}, \boldsymbol{I}_L]^T, \boldsymbol{I}_N^T]^T$ with $\boldsymbol{I}_l$ being the identity matrix of size l, and $\boldsymbol{b} := [b(0), \ldots, b(N-1)]^T$ is an $N \times 1$ vector.

Discarding the CP from $\bar{\boldsymbol{x}}(i)$ leads to

$$\boldsymbol{x}(i) = \boldsymbol{H}\boldsymbol{u}(i) + \boldsymbol{w}(i), \tag{3}$$

where $\boldsymbol{u}(i) = \boldsymbol{A}\boldsymbol{s}(i) + \boldsymbol{b}$; $\boldsymbol{h} := [h(0), h(1), \ldots, h(L)]^T$; and $\boldsymbol{H}$ is an $N \times N$ circulant matrix with first column $[\boldsymbol{h}^T, 0, \ldots, 0]^T$. We will further make the following assumptions.

A1. The information symbol vector is zero-mean, stationary, with correlation matrix $E\{\boldsymbol{s}(i)\boldsymbol{s}^{\mathcal{H}}(i)\} := \boldsymbol{R}_s > \boldsymbol{0}$, where $^{\mathcal{H}}$ denotes conjugated transposition.

A2. The noise $\boldsymbol{w}(i)$ is zero-mean, Gaussian, with correlation matrix $\sigma_w^2\boldsymbol{I}$.

We model the channel to be a stochastic process in the following way.

A3. Each channel tap $h(l)$ is complex Gaussian distributed with zero-mean and variance $\sigma_{h_l}^2$, denoted by $\mathcal{CN}(0, \sigma_{h_l}^2)$. Channel taps are independent of each other.

A3 is well justified especially for wireless fading channels that are rich in scattering. Notice that channel taps are not necessarily identically distributed; i.e., A3 allows for taps with, e.g., an exponential power profile.

Let $\boldsymbol{F}$ be the $N \times N$ discrete fourier transform (DFT) matrix with (m,n)th entry $[\boldsymbol{F}]_{m,n} = N^{-\frac{1}{2}}W^{-mn}$, where $W := \exp(j2\pi/N)$. Define the channel transfer function as $H(z) := \sum_{n=0}^{L} h(n)z^{-n}$. We can express the channel frequency response in vector form as

$$[H(1), H(W), \ldots, H(W^{N-1})]^T = N^{\frac{1}{2}}\boldsymbol{F}_{\bar{L}}\boldsymbol{h}$$

where $\boldsymbol{F}_{\bar{L}}$ is a submatrix of $\boldsymbol{F}$, corresponding to the first $\bar{L}$ columns of $\boldsymbol{F}$. It follows from A3 that $H(W^n)$ for any $n \in [0, N-1]$ is $\mathcal{CN}(0, \sigma_H^2)$, where

$$\sigma_H^2 := \sum_{l=0}^{L} \sigma_{h_l}^2. \tag{4}$$

This implies that the channel frequency response samples $\{H(W^n)\}_{n\in[0,N-1]}$ are identically distributed, Gaussian, with zero mean, and variance σ_H^2. Notice that even though the channel taps do not have the same variance, its frequency response samples at the DFT grid have the same variance.

Since the channel varies from block to block and is unknown to the receiver, we have to estimate it by using only one received block. As we consider block-by-block processing, we omit the block indices in (3), and re-express it as

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{B}\boldsymbol{h} + \boldsymbol{w} \tag{5}$$

where $\boldsymbol{B}$ is an $N \times \bar{L}$ column-wise circulant matrix with first column $\boldsymbol{b}$; and in deriving (5) we used the commutativity of circular convolution to obtain $\boldsymbol{H}\boldsymbol{b} = \boldsymbol{B}\boldsymbol{h}$.

The model (5) can be viewed as a *virtual* two-user model where one user transmits $\boldsymbol{s}$ and the other one $\boldsymbol{h}$ through equivalent channels $\boldsymbol{H}\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. They interfere with each other, and thus it is desirable to decouple them, which in our single-user model amounts to separating channel estimation from symbol recovery. We have established in [20] that in the absence of noise, $\boldsymbol{s}$ and $\boldsymbol{h}$ can be decoupled for any FIR channel of order L if $\boldsymbol{B}$ has full column rank, and the following condition is satisfied for $\forall\, \boldsymbol{H}$:

$$\boldsymbol{B}^{\mathcal{H}}\boldsymbol{H}\boldsymbol{A} = \boldsymbol{0}. \tag{6}$$

Notice that if (6) holds true, then

$$\boldsymbol{B}^{\mathcal{H}}\boldsymbol{x} = \boldsymbol{B}^{\mathcal{H}}\boldsymbol{H}\boldsymbol{A}\boldsymbol{s} + \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}\boldsymbol{h} = \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}\boldsymbol{h}$$

which implies that one can isolate $\boldsymbol{h}$.

Expressing the DFT of the training blocks as $\tilde{\boldsymbol{b}} := \boldsymbol{F}\boldsymbol{b}$, and denoting its i_kth entry as $\tilde{b}_{i_k}$, let us define the set of K *ordered* integer indices

$$\mathcal{I}^{\perp} := \{i_k | \tilde{b}_{i_k} \neq 0, i_k < i_{k+1}, k \in [0, K-1]\} \tag{7}$$

and its complement $\mathcal{I}$ containing $M = N - K$ indices i for which $\tilde{b}_i = 0$. If we define the $K \times 1$ vector $\check{\boldsymbol{b}}$ to contain the K nonzero entries of the DFT pilot vector $\tilde{\boldsymbol{b}}$ so that $\check{b}_k = \tilde{b}_{i_k}$, then the time-domain training vector can be written as

$$\begin{aligned}\boldsymbol{b} = \boldsymbol{F}^{\mathcal{H}}\tilde{\boldsymbol{b}} &= \boldsymbol{F}^{\mathcal{H}}\boldsymbol{P}_{\mathcal{I}}\begin{bmatrix}\check{\boldsymbol{b}}\\ \boldsymbol{0}\end{bmatrix}\\ &= \boldsymbol{F}^{\mathcal{H}}\boldsymbol{P}_{\mathcal{I}^{\perp}}\begin{bmatrix}\boldsymbol{0}\\ \check{\boldsymbol{b}}\end{bmatrix}\end{aligned} \tag{8}$$

where $\boldsymbol{P}_{\mathcal{I}}(\boldsymbol{P}_{\mathcal{I}^{\perp}})$ is a permutation matrix collecting the K possibly dispersed nonzero entries of $\tilde{\boldsymbol{b}}$ at the top (bottom). Correspondingly, similar to [20], we focus on linear precoding matrices that can be factored as

$$\boldsymbol{A} = \boldsymbol{F}^{\mathcal{H}}\boldsymbol{P}_{\mathcal{I}^{\perp}}\begin{bmatrix}\boldsymbol{I}_M\\ \boldsymbol{0}_{K\times M}\end{bmatrix}. \tag{9}$$

With this transmitted block $\boldsymbol{u} = \boldsymbol{F}^{\mathcal{H}}\boldsymbol{P}_{\mathcal{I}^{\perp}}[\boldsymbol{s}^T, \check{\boldsymbol{b}}^T]^T$, the symbols $\boldsymbol{s}$ are loaded on M subcarriers (the columns of $\boldsymbol{F}^{\mathcal{H}}$ in (9)) that are distinct from the K subcarriers which correspond to the pilots used for channel estimation. Since $\boldsymbol{s}$ and $\check{\boldsymbol{b}}$ are loaded separately in the frequency domain, it is easy to check that indeed $(\boldsymbol{A}, \boldsymbol{b})$ satisfies (6).

For *deterministic* channels, $(\boldsymbol{A}, \boldsymbol{b})$ was designed in [20] to enable both optimal least squares (LS) channel estimation and minimization of the symbol MSE. In this correspondence, we consider the stochastic counterpart of [20], and design pilots under the following design constraint.

C2. The pair $(\boldsymbol{A}, \boldsymbol{b})$ is selected from the class that satisfies (8) and (9).

Since (8) and (9) suggest frequency-division multiple access (FDMA)-like transmission of information symbols and pilots, we will find it convenient to separate the received block in the frequency domain. Toward this objective, recall that the circulant channel matrix

$\boldsymbol{H}$ can be diagonalized by the DFT matrix $\boldsymbol{F}$ and its inverse (IDFT) [9, p. 202], such that

$$\boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^{\mathcal{H}} = \text{diag}[H(W^0), H(W^1), \ldots, H(W^{N-1})]. \tag{10}$$

Consider the DFT of the IBI-free received block, expressed as $\tilde{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x}$, and premultiply $\tilde{\boldsymbol{x}}$ by $\boldsymbol{P}_{\mathcal{I}^\perp}$, to rearrange its entries. Substituting (8) and (9) into (5), and taking into account (9) and (10), we can express $\boldsymbol{P}_{\mathcal{I}^\perp}\boldsymbol{F}\boldsymbol{x}$ as

$$\begin{aligned}\boldsymbol{P}_{\mathcal{I}^\perp}\boldsymbol{F}\boldsymbol{x} &:= \begin{bmatrix}\tilde{\boldsymbol{x}}_s\\ \tilde{\boldsymbol{x}}_b\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{D}_{H,s} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{D}_{H,b}\end{bmatrix}\begin{bmatrix}\boldsymbol{s}\\ \check{\boldsymbol{b}}\end{bmatrix} + \begin{bmatrix}\tilde{\boldsymbol{w}}_s\\ \tilde{\boldsymbol{w}}_b\end{bmatrix}\end{aligned} \tag{11}$$

where $\boldsymbol{D}_{H,s}$ and $\boldsymbol{D}_{H,b}$ are diagonal matrices satisfying

$$\text{diag}(\boldsymbol{D}_{H,s}, \boldsymbol{D}_{H,b}) = \boldsymbol{P}_{\mathcal{I}^\perp}\boldsymbol{F}\boldsymbol{H}\boldsymbol{F}^{\mathcal{H}}\boldsymbol{P}_{\mathcal{I}^\perp}$$

and $\tilde{\boldsymbol{w}}_s(\tilde{\boldsymbol{w}}_b)$ denotes the DFT of the noise $\tilde{\boldsymbol{w}}$ corresponding to the information (training) part.

Relying on the decoupled subblocks $\tilde{\boldsymbol{x}}_s$ and $\tilde{\boldsymbol{x}}_b$ in (11), we will first form the channel estimate $\hat{\boldsymbol{h}}$ based on $\tilde{\boldsymbol{x}}_b$, and then recover the channel frequency response as

$$[\hat{H}(W^{i_0}), \hat{H}(W^{i_1}), \ldots, \hat{H}(W^{i_{M-1}})]^T = N^{\frac{1}{2}}\check{\boldsymbol{F}}_{\bar{L}}\hat{\boldsymbol{h}} \tag{12}$$

where $\check{\boldsymbol{F}}_{\bar{L}}$ is an $M \times \bar{L}$ matrix with nth row equal to the i_nth row of $\boldsymbol{F}_{\bar{L}}$ for $i_n \in \mathcal{I}$. Based on the estimated channel frequency response, we can subsequently equalize $\tilde{\boldsymbol{x}}_s$ in (11) to recover the information symbols $\boldsymbol{s}$.

III. Optimal Training for LMMSE Channel Estimation

The placement and power allocation of pilot symbols will be sought in this section to minimize the channel MSE subject to a fixed transmit-power constraint. Equispaced and equipowered pilots will turn out to be the optimal choice.

Let $B(z)$ be the $\mathcal{Z}$-transform of $\boldsymbol{b}$ defined as

$$B(z) := \sum_{n=0}^{N-1} b(n)z^{-n}$$

and $\boldsymbol{D}_B$ a diagonal matrix defined as

$$\boldsymbol{D}_B := \text{diag}[B(1), B(W), \ldots, B(W^{N-1})].$$

Since $\boldsymbol{B}$ is a tall and column-wise circulant matrix, its DFT-based diagonalization yields

$$\boldsymbol{B} = \boldsymbol{F}^{\mathcal{H}}\boldsymbol{D}_B\boldsymbol{F}_{\bar{L}}. \tag{13}$$

For simplicity, we temporarily assume that $\boldsymbol{A} = \boldsymbol{0}$ in (5), i.e., $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{h} + \boldsymbol{w}$. Since this is a linear regression model with $\boldsymbol{h}$ and $\boldsymbol{w}$ complex normal and $\boldsymbol{B}$ known, the linear MMSE (LMMSE) estimator for $\boldsymbol{h}$ coincides with the MMSE channel estimator.

Let us define the channel estimation error as $\Delta\boldsymbol{h} := \boldsymbol{h} - \hat{\boldsymbol{h}}$. For a given $\boldsymbol{B}$, the LMMSE channel estimator and the corresponding MMSE are given by [14, Ch. 12]

$$\hat{\boldsymbol{h}} := \boldsymbol{R}_{hx}\boldsymbol{R}_x^{-1}\boldsymbol{x} \tag{14}$$

$$\sigma^2_{\Delta h} := \text{trace}(\boldsymbol{\Sigma}) = \text{trace}\left(\boldsymbol{R}_h - \boldsymbol{R}_{hx}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{hx}^{\mathcal{H}}\right) \tag{15}$$

where $\boldsymbol{R}_x = E\{\boldsymbol{x}\boldsymbol{x}^{\mathcal{H}}\}$, $\boldsymbol{R}_{hx} = E\{\boldsymbol{h}\boldsymbol{x}^{\mathcal{H}}\}$, $\boldsymbol{R}_h = E\{\boldsymbol{h}\boldsymbol{h}^{\mathcal{H}}\}$, and

$$\boldsymbol{\Sigma} := E\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\mathcal{H}}\}. \tag{16}$$

We substitute $\boldsymbol{R}_x = \boldsymbol{B}\boldsymbol{R}_h\boldsymbol{B}^{\mathcal{H}} + \sigma_w^2\boldsymbol{I}$ and $\boldsymbol{R}_{hx} = \boldsymbol{R}_h\boldsymbol{B}^{\mathcal{H}}$ into (15), and apply the matrix inversion lemma to obtain

$$\sigma^2_{\Delta h} = \text{trace}\left[\left(\boldsymbol{R}_h^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}\right)^{-1}\right]. \tag{17}$$

Also from the matrix inversion lemma, we can express $\hat{\boldsymbol{h}}$ as

$$\hat{\boldsymbol{h}} := \frac{1}{\sigma_w^2}\left(\boldsymbol{R}_h^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{x}. \tag{18}$$

From (18) and (6), we find that even for $\boldsymbol{A} \neq \boldsymbol{0}$

$$\begin{aligned}&(1/\sigma_w^2)\left(\boldsymbol{R}_h^{-1} + \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}/\sigma_w^2\right)^{-1}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{x}\\ &\qquad= (1/\sigma_w^2)(\boldsymbol{R}_h^{-1} + \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}/\sigma_w^2)^{-1}\boldsymbol{B}^{\mathcal{H}}(\boldsymbol{B}\boldsymbol{h} + \boldsymbol{w}).\end{aligned}$$

This shows that the information symbol block does not affect the channel estimator, thanks to the decoupling imposed by (6). Thus, we can conclude that $\hat{\boldsymbol{h}}$ is the channel MMSE estimator even if $\boldsymbol{A} \neq \boldsymbol{0}$.

We also have from (11) and (13) that

$$\begin{aligned}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{x} &= \boldsymbol{F}_{\bar{L}}^{\mathcal{H}}\boldsymbol{D}_B^{\mathcal{H}}\boldsymbol{F}\boldsymbol{x} = \boldsymbol{F}_{\bar{L}}^{\mathcal{H}}\boldsymbol{D}_B^{\mathcal{H}}\boldsymbol{P}_{\mathcal{I}^\perp}\begin{bmatrix}\tilde{\boldsymbol{x}}_s\\ \tilde{\boldsymbol{x}}_b\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{0} & \check{\boldsymbol{F}}_{\bar{L}}^{\mathcal{H}}\check{\boldsymbol{D}}_B^{\mathcal{H}}\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{x}}_s\\ \tilde{\boldsymbol{x}}_b\end{bmatrix} = \check{\boldsymbol{F}}_{\bar{L}}^{\mathcal{H}}\check{\boldsymbol{D}}_B^{\mathcal{H}}\tilde{\boldsymbol{x}}_b\end{aligned} \tag{19}$$

where $\check{\boldsymbol{D}}_B$ denotes the $K \times K$ submatrix of $\boldsymbol{D}_B$ with all nonzero diagonal entries, and $\check{\boldsymbol{F}}_{\bar{L}}$ is the $K \times \bar{L}$ matrix formed by the K rows of $\boldsymbol{F}_{\bar{L}}$ corresponding to the nonzero entries of equation $\boldsymbol{D}_B$. Equation (19) shows that the MMSE channel estimator in (18) requires only equation $\tilde{\boldsymbol{x}}_b$; hence, $\hat{\boldsymbol{h}}$ can be obtained via training based on pilot tones.

Now, we will design $\boldsymbol{B}$ (or, equivalently, $\boldsymbol{b}$) to minimize $\sigma^2_{\Delta h}$ subject to the power constraint $\text{trace}(\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}) = \bar{L}\|\boldsymbol{b}\|^2 = \bar{L}\mathcal{P}_b$, i.e., we will find the optimal $\boldsymbol{B}$ as

$$\boldsymbol{B}_{opt} = \arg_{\boldsymbol{B}} \min_{\text{trace } \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B} = \bar{L}\mathcal{P}_b} \text{trace}\left[\left(\boldsymbol{R}_h^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}\right)^{-1}\right]. \tag{20}$$

In Appendix I, we show that the channel MSE is lower-bounded as follows:

$$\begin{aligned}\text{trace}\left[\left(\boldsymbol{R}_h^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}\right)^{-1}\right] &\geq \sum_{l=0}^{L}\frac{1}{\left[\boldsymbol{R}_h^{-1} + \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}/\sigma_w^2\right]_{l,l}}\\ &= \sum_{l=0}^{L}\frac{\sigma_{h_l}^2\sigma_w^2}{\sigma_w^2 + \mathcal{P}_b\sigma_{h_l}^2}\end{aligned} \tag{21}$$

where the equality holds if and only if $\boldsymbol{R}_h^{-1} + \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}/\sigma_w^2$ is diagonal. Since $\boldsymbol{R}_h^{-1}$ is diagonal, and every diagonal entry of $\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}$ is $\mathcal{P}_b$, we infer that $\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}$ has to be $\mathcal{P}_b\boldsymbol{I}$. It readily follows from (13) that $\boldsymbol{B}^{\mathcal{H}}\boldsymbol{B} = \mathcal{P}_b\boldsymbol{I}$ if $K(\geq \bar{L})$ pilot tones are equispaced and equipowered, i.e., $|B(W^i)|^2 = N\mathcal{P}_b/K$ for $i \in \mathcal{I}^\perp$. Moreover, for $K = \bar{L}$, we can show as in [20] that only equipowered and equispaced pilot tones achieve this minimum channel MSE.

Notice that we did not impose any restriction on $\boldsymbol{b}$ except for constraining its power. This implies that among all training possibilities, usage of equispaced and equipowered pilot tones is optimal in the MMSE sense. In summary, we have established the following.

Theorem 1: Suppose that C1 and C2 hold true. Under A1-A4, the minimum channel MSE for a fixed power constraint $\|\boldsymbol{b}\|^2 = \mathcal{P}_b$ is always attained by K equispaced and equipowered pilot tones, provided that $K \geq \bar{L}$. The minimum channel MSE is given by (21).

Prompted by Theorem 1, we will henceforth focus on equispaced pilot tones that we formally introduce as

C3. The block size N is a multiple of K such that $N = KJ$ for $K \geq \bar{L}$, and for some positive integer J; the pilots are chosen to be equispaced and equipowered with the index set in (7) as $\mathcal{I}^{\perp} := \{j + Jk | k \in [0, K-1]\}$ for some $j \in [0, J-1]$.

Condition C3 implies that the nonzero entries of the DFT of $\boldsymbol{b}$ are equispaced. When $j \neq 0$, equispaced should be understood in a circular sense; i.e., after periodically repeating $\tilde{\boldsymbol{b}}$. It should be noted that Theorem 1 only requires $K \geq \bar{L}$ and says nothing on: a) whether L is the minimum number of pilots needed from a capacity perspective; and b) how the overall power should be allocated to training and information-bearing symbols, which we will optimize in terms of the capacity bound in the Section IV

IV. Designing Pilots Based on a Capacity-Bound Criterion

We have developed the optimal MMSE pilot tones. But, there still exist degrees of freedom in the number of pilots and the power allocation between pilot and information-bearing symbols. To enhance the capacity of our system, we further design the number of pilots and the power allocation, using a capacity-bound criterion.

Suppose first that the channel estimation step is perfect, and let the transmit power be $\mathcal{P} := \mathcal{P}_s + \mathcal{P}_b$ with $\mathcal{P}_s := E\{\|\boldsymbol{s}\|^2\}$ and $\mathcal{P}_b = \|\boldsymbol{b}\|^2$. Under this transmit power constraint, and following standard steps (as in [5], [7], [17], [26]), we have that the channel capacity (normalized per transmitted symbol) averaged over the random matrix $\boldsymbol{D}_{H,s}$ is given by

$$C_{\text{ideal}} = \frac{M}{\bar{N}} E\left\{\log\left(1 + \rho_{\text{ideal}} |g|^2\right)\right\} \tag{22}$$

where the expectation is taken with respect to $g \sim \mathcal{CN}(0,1)$, and ρ_{ideal} denotes the signal-to-noise ratio (SNR) that is defined as

$$\rho_{\text{ideal}} := \frac{\sigma_H^2 \mathcal{P}_s}{M \sigma_w^2}. \tag{23}$$

Under the assumption that the channel estimation step is perfect, it is easy to see that C_{ideal} will be maximized if we set all transmitted symbols to be information-bearing symbols, i.e., $M = N$ and $\mathcal{P} = \mathcal{P}_s$. This confirms that C_{ideal} does not account for channel estimation accuracy. We will incorporate the channel estimation error of equispaced and equipowered pilots into the average channel capacity to find the minimum number of pilots needed and the overall power allocated between training and information-bearing symbols.

Applying on a per-subcarrier basis the lower bound on average channel capacity that is derived in [18, eq. (46)] and [5, eq. 3.3.55] for serial transmissions over flat-fading channels, and summing across subcarriers, we obtain the following:

$$C \geq \underline{C} := \frac{1}{\bar{N}} \sum_{i \in \mathcal{I}} E\left\{\log\left(1 + \frac{|\hat{H}(W^i)|^2}{E\{|\Delta H(W^i)|^2\} + M\sigma_w^2/\mathcal{P}_s}\right)\right\} \tag{24}$$

where $\Delta H(W^i) = H(W^i) - \hat{H}(W^i)$. For MMSE estimators, the lower bound $\underline{C}$ coincides with the GMI, which is the maximum rate at which the average probability of error converges to zero as the codeword length increases [16]. Although Gaussian codes of infinite block size (and, hence, infinite $\bar{N}$) are required in theory to attain the channel capacity, we will demonstrate by simulations that the results obtained from our capacity analysis are beneficial even for a finite number of symbols drawn from a finite alphabet.

Suppose that we fix the transmit power for pilot tones, and for information-bearing symbols. We will find it convenient to work with the normalized (unit-variance) LMMSE channel estimator $g := \hat{H}(W^i)/E^{1/2}\{|\hat{H}(W^i)|^2\}$; note that g is not a function of i since the normalized variates are identically distributed $\forall\ i$. Using the orthogonality principle, we can express $\underline{C}$ in (24) as

$$\begin{aligned}\underline{C} &= \frac{1}{\bar{N}} \sum_{i \in \mathcal{I}} E\left\{\log\left(1 + \frac{E\{|\hat{H}(W^i)|^2\}|g|^2}{E\{|\Delta H(W^i)|^2\} + M\sigma_w^2/\mathcal{P}_s}\right)\right\} \\ &= \frac{1}{\bar{N}} \sum_{i \in \mathcal{I}} E\left\{\log\left(1 + \frac{(\sigma_H^2 - E\{|\Delta H(W^i)|^2\})|g|^2}{E\{|\Delta H(W^i)|^2\} + M\sigma_w^2/\mathcal{P}_s}\right)\right\}. \end{aligned} \tag{25}$$

Under C3, we can verify that $E\{|\Delta H(W^i)|^2\} = \sigma_{\Delta h}^2\ \forall\ i \in \mathcal{I}$ (see Appendix B). This allows us to rewrite (25) as

$$\underline{C} = \frac{M}{\bar{N}} E\left\{\log\left(1 + \rho|g|^2\right)\right\} \tag{26}$$

where

$$\rho := \frac{\sigma_H^2 - \sigma_{\Delta h}^2}{\sigma_{\Delta h}^2 + M\sigma_w^2/\mathcal{P}_s}. \tag{27}$$

A. Optimal Number of Pilots

Clearly, the number of pilots affects bandwidth efficiency that depends on the relative redundancy K/M, and is defined as

$$\mathcal{E}(M, K) := M/\bar{N} = M/(M + K + L).$$

For a fixed M, the minimum K, i.e., $K = \bar{L}$, is optimal in terms of bandwidth efficiency. The minimum K also minimizes complexity and decoding delay at the receiver end. We will show that the minimum K is also optimal in the sense of maximizing the lower bound on the average channel capacity.

Suppose that the power $\mathcal{P}_b$ alloted for channel estimation is constant, and that we fix the block size N. To obtain the optimal M (or, equivalently, K since $M = N - K$), we re-express M as μ, and treat μ as a continuous variable, assuming that any μ achieves the minimum channel MSE. We differentiate $\underline{C}$ in (26) with respect to μ to obtain

$$\bar{N}\frac{\partial \underline{C}}{\partial \mu} = E\left\{\log\left(1 + \rho|g|^2\right) + \mu \frac{\partial \rho}{\partial \mu} \frac{|g|^2}{1 + \rho|g|^2}\right\}. \tag{28}$$

For a given $\mathcal{P}_b$, the variance $\sigma_{\Delta h}^2$ is found to be independent of μ. It follows from (27) that

$$\frac{\partial \rho}{\partial \mu} = -\frac{\left(\sigma_H^2 - \sigma_{\Delta h}^2\right)\sigma_w^2/\mathcal{P}_s}{\left(\sigma_{\Delta h}^2 + \mu\sigma_w^2/\mathcal{P}_s\right)^2} = -\rho \frac{1}{\sigma_{\Delta h}^2/\left(\sigma_w^2/\mathcal{P}_s\right) + \mu}. \tag{29}$$

Substituting (29) into (28), we arrive at

$$\begin{aligned}\bar{N}\frac{\partial \underline{C}}{\partial \mu} &= E\left\{\log\left(1 + \rho|g|^2\right) - \frac{\rho|g|^2}{1 + \rho|g|^2} \frac{\mu}{\sigma_{\Delta h}^2/(\sigma_w^2/\mathcal{P}_s) + \mu}\right\} \\ &\geq E\left\{\frac{\rho|g|^2}{1 + \rho|g|^2}\left(1 - \frac{\mu}{\sigma_{\Delta h}^2/(\sigma_w^2/\mathcal{P}_s) + \mu}\right)\right\} \\ &\geq E\left\{\frac{\rho|g|^2}{1 + \rho|g|^2} \frac{\sigma_{\Delta h}^2/(\sigma_w^2/\mathcal{P}_s)}{\sigma_{\Delta h}^2/(\sigma_w^2/\mathcal{P}_s) + \mu}\right\} > 0 \end{aligned} \tag{30}$$

where in the second step, we used the inequality $\log(1 + \rho|g|^2) \geq \rho|g|^2/(1+\rho|g|^2)$. Since $\partial\underline{C}/\partial\mu > 0$, to achieve the minimum channel MSE, we should take μ as large as possible; that is, $\mu = N - \bar{L}$, or equivalently, $K = \bar{L}$. Moreover, for $K = \bar{L}$, only equipowered and equispaced pilot tones achieve the minimum channel MSE [20, Theorem 1].

We note that our result regarding optimality of the minimum number of pilots, independent of the allocated power for training, differs markedly from the corresponding result reported in [10, Theorem 3] for multiple-input multiple-output (MIMO) channels. The reason is that we first optimize the number of pilots and then the allocated power, while [10] optimizes first the allocated power and then the number of pilots. We conjecture (and it will be interesting to prove) that the minimum number of pilots for MIMO channels is independent of the allocated training power, but this goes beyond the scope of this correspondence.

B. Optimal Power Allocation

The last question we wish to address pertains to the optimal distribution of the transmitted power between information and training symbols. We will denote the power allocated to information-bearing symbols and pilot symbols, respectively, as

$$E\{\|\boldsymbol{s}\|^2\} = \mathcal{P}_s = \alpha\mathcal{P}, \qquad \|\boldsymbol{b}\|^2 = \mathcal{P}_b = (1-\alpha)\mathcal{P} \tag{31}$$

for $0 < \alpha < 1$, where α represents the percentage of power allocated to information symbols. Plugging (M, L) into (26), $\underline{C}$ becomes a function of α only. Thus, we can find the optimal α numerically. When the MMSE $\sigma^2_{\Delta h}$ can be explicitly expressed as a function α, closed-form expressions for the optimal α become available. We discuss two such cases next for a fixed (M, K) pair with $K \geq L$.

1) High-SNR Regime: At high SNR, we have that

$$(\sigma^2_H - \sigma^2_{\Delta h}) \cong \sigma^2_H \quad \text{and} \quad \sigma^2_{\Delta h} \cong \bar{L}\sigma^2_w/\mathcal{P}_b.$$

From $\mathcal{P}_s = \alpha\mathcal{P}$ and $\mathcal{P}_b = (1-\alpha)\mathcal{P}$, we obtain

$$\rho = \frac{\sigma^2_H}{\bar{L}\sigma^2_w/\mathcal{P}_b + M\sigma^2_w/\mathcal{P}_s} = M\rho_{\rm snr}\frac{\alpha(1-\alpha)}{\bar{L}\alpha + M(1-\alpha)} \tag{32}$$

where $\rho_{\rm snr}$ represents the output SNR defined as

$$\rho_{\rm snr} = \frac{\sigma^2_H\mathcal{P}}{M\sigma^2_w}. \tag{33}$$

For a fixed pair (M, K), the ratio M/N is constant. Since $\log(\cdot)$ is an increasing function, maximizing $\underline{C}$ with respect to α is equivalent to maximizing ρ. Differentiating ρ with respect to α, we find that ρ is maximized at

$$\alpha_\infty := \frac{1}{1+\sqrt{\bar{L}/M}} \tag{34}$$

and its maximum is given by

$$\rho_\infty := \rho_{\rm snr}\alpha^2_\infty. \tag{35}$$

Interestingly, α_∞ coincides with the power allocation ratio in [20], which was obtained by minimizing the symbol MSE assuming an underlying *deterministic* channel estimated based on the LS criterion.

For a fixed power of the information-bearing symbols, setting the estimation error variance in $\underline{C}$ in (26) to be zero, we obtain the *upper bound* of $\underline{C}$, which is the ideal average capacity without channel estimation error. In this case, ρ equals to $\rho_{\rm ideal}$ defined by (23). We can define the ratio $\rho_{\rm snr}/\rho_{\rm ideal}$ to measure the performance loss due to channel estimation. In this case, from $\rho_{\rm ideal} = \alpha_\infty\rho_{\rm snr}$, we find $\rho_{\rm snr}/\rho_{\rm ideal} = \alpha_\infty$. Then, it follows from (34) that $\alpha_\infty \geq 0.5$. We thus deduce that optimally powered pilots incur an asymptotic (high-SNR) performance loss of at most $10\log(0.5) \approx 3.0$ dB, due to the channel estimation error. And this asymptotic performance loss converges to zero as the ratio $\bar{L}/M$ decreases between the channel length and the information-bearing block size increases. These observations also imply that the lower bound is tight for $\bar{L}/M \to 0$ at high SNR.

2) Rayleigh Channels With equiPowered Taps: Here we consider that the channel taps are i.i.d., zero-mean, complex Gaussian $\mathcal{CN}(0, \sigma^2_h)$. In this case, from (21), we obtain

$$\sigma^2_{\Delta h} = \bar{L}\sigma^2_h\sigma^2_w/(\sigma^2_w + \mathcal{P}_b\sigma^2_h)$$

and

$$\sigma^2_H - \sigma^2_{\Delta h} = \bar{L}\mathcal{P}_b\sigma^4_h/(\sigma^2_w + \mathcal{P}_b\sigma^2_h).$$

We show in Appendix III that for $M > \bar{L}$, the SNR ρ can be re-expressed as

$$\rho = \frac{M\rho_{\rm snr}}{M-\bar{L}}\frac{\alpha(1-\alpha)}{(\beta-\alpha)} \tag{36}$$

where

$$\beta := \frac{M}{M-\bar{L}}\left(1+\frac{\bar{L}}{M\rho_{\rm snr}}\right). \tag{37}$$

Differentiating $f(\alpha) := \alpha(1-\alpha)/(\beta-\alpha)$ yields $f'(\alpha) = (\alpha^2 - 2\beta\alpha + \beta)/(\beta-\alpha)^2$, from which we find that ρ is maximized at

$$\alpha_{\rm iid} := \frac{1}{1+\sqrt{1-1/\beta}}. \tag{38}$$

Substituting this into ρ leads to

$$\rho_{\rm iid} := \frac{\rho_{\rm snr}}{1+\bar{L}/(M\rho_{\rm snr})}\alpha^2_{\rm iid}. \tag{39}$$

The performance loss $\rho_{\rm iid}/\rho_{\rm ideal}$ due to the channel estimation error is given by $\alpha_{\rm iid}/[1+\bar{L}/(M\rho_{\rm snr})]$. Since β is a decreasing function in SNR, $\alpha_{\rm iid}$ is found from (38) to be an increasing function in SNR, and is bounded such that $\alpha_{\rm iid} \geq 0.5$. At $\rho_{\rm snr} = \bar{L}/M$, we have $\rho_{\rm iid}/\rho_{\rm ideal} \geq 0.25$. Thus, for any M, the performance loss due to channel estimation is no more than about 6.0 dB at $\rho_{\rm snr} = \bar{L}/M$, and converges to 3.0 dB at high SNR.

V. Numerical Examples

To validate our analysis and design, we test uncoded OFDM transmissions with $N = 64$ subcarriers, and CP length (channel order) equal to $L = 7$. We also insert $\bar{L} = L + 1 = 8$ equispaced and equipowered pilot tones. We deal with two cases: i) channels having i.i.d. complex zero-mean Gaussian taps; and ii) channels with complex zero-mean Gaussian taps, but with exponential power profile such that $\sigma^2_{h_l} = \exp(-l)$ for $l \in [0, \bar{L}]$.

A. Average Channel Capacity

For channel taps with identical power profiles, Fig. 1 compares the average channel capacity when: i) the channel is known to the receiver and training is not required, i.e., $M = N$ and $\alpha = 1$; ii) the upper and lower bounds on the average capacity for the optimal power allocation, where the optimal α is computed by (38); and iii) the fixed $\alpha = 1/(1+\bar{L}/M) = 0.875$, where all symbols have the same power as considered in [6], [19].

The difference between the average channel capacity without training and the upper bound can be considered as the price we have to pay to acquire CSI. The difference between the lower and the upper bound is the penalty resulting from the worst channel estimation error. The optimally powered pilots have at most

$$-10\log\alpha_{\rm iid}/[1+\bar{L}/(M\rho_{\rm snr})] \approx -10\log 0.72 \approx 1.43 \text{ dB}$$

performance loss at 10 dB due to the channel estimation error, which converges to

$$-10\log\alpha_\infty = -10\log 1/(1+\sqrt{1/7}) \approx 1.39 \text{ dB}$$

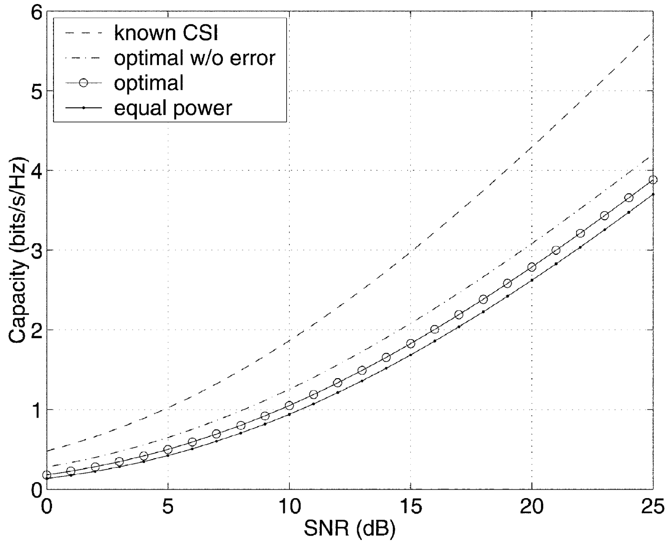

Fig. 1. Channel capacity bounds versus SNR (identical power profile).

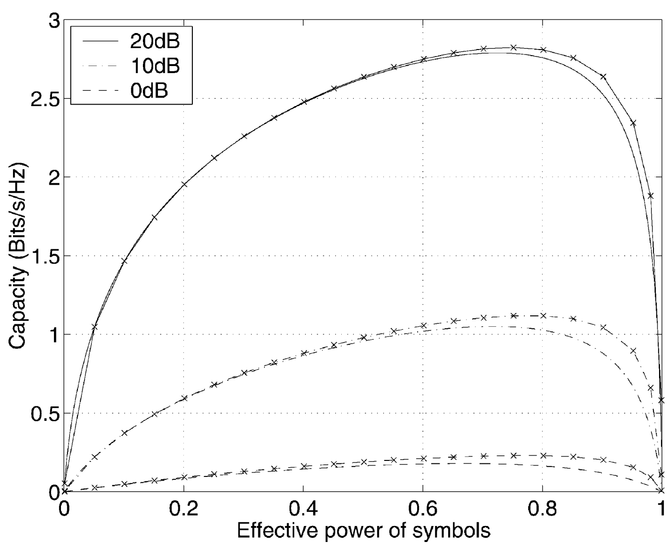

Fig. 2. Channel capacity bounds versus α: identical power profile; exponential power profile (with $\times$).

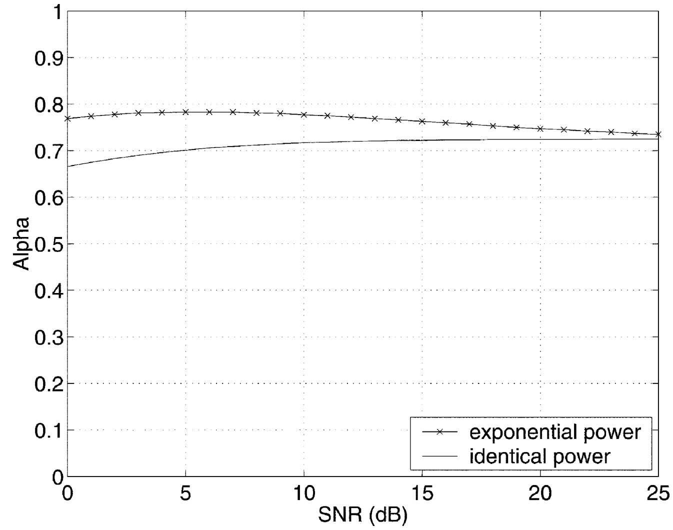

Fig. 3. Optimal α versus SNR.

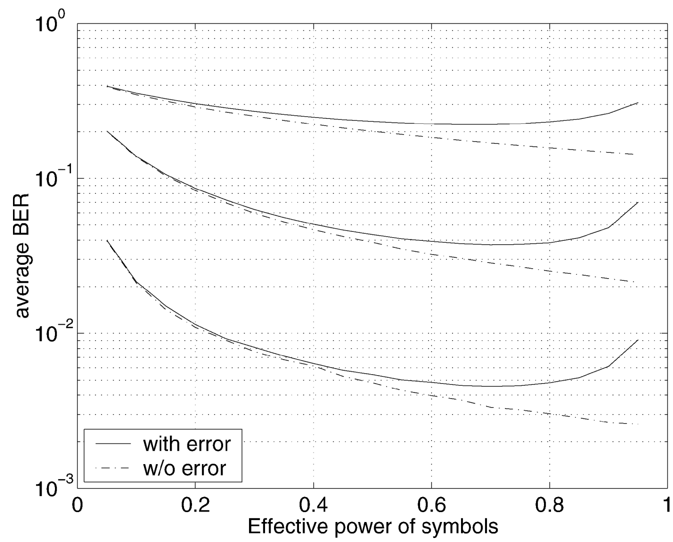

Fig. 4. Average BER versus α (i.i.d. channel taps).

as SNR increases. Recall that the actual average channel capacity lies between the lower and the upper bound. Their small difference implies the tightness of the lower bound, and hence validates our design of training pilots.

Optimally powered pilots gain about 1 dB against the fixed powered pilots. Even at low SNR, there is a small difference between them. This is because the lower bound is a flat function around its maximum as illustrated in Fig. 2. It can also be observed that although channels with exponential power profiles can be quite different from those with identical power profiles, their capacity bounds are almost the same. Fig. 3 depicts the optimal α as a function of SNR. As expected by theory, for both cases, the optimal α converges to α_∞, as SNR increases.

B. BER Performance

Although BER performance for a specific constellation may not relate directly to average channel capacity, we compare BER of the transmission schemes described above to justify our optimal design. We computed MMSE channel estimators based on pilot tones, from which we obtained maximum-likelihood (ML) symbol estimates. In every simulation, 1000 channel realizations were used, and the results were averaged. The results for channels with exponential power profiles are omitted because they are quite similar with those channels having identical power profiles.

Fig. 4 illustrates the BER as a function of the effective power α of information symbols. The dashed-dotted curves correspond to the BER without estimation errors, i.e., when channel estimation is perfect. When α is small, i.e., when we allocate much power for our pilots to yield accurate channel estimates, the performance loss due to the channel estimation error is small. As α goes to 1, the difference diverges because without channel estimation error, the BER becomes an increasing function of α. It is observed that the minimum BER is attained around $\alpha = 0.7$, which almost equals the $\alpha_{\rm iid}$ suggested by the lower bound on the average channel capacity.

Fig. 5 reports the average BER as a function of SNR. Comparing with the BER of the known CSI case, we infer that the total cost to acquire CSI with optimally powered pilots is about 1.5 dB. The penalty of inaccurate CSI in BER is about 1.5 dB, and the optimally powered pilots have a slight performance gain over the fixed-power pilots. These are quite similar conclusions with those obtained with average

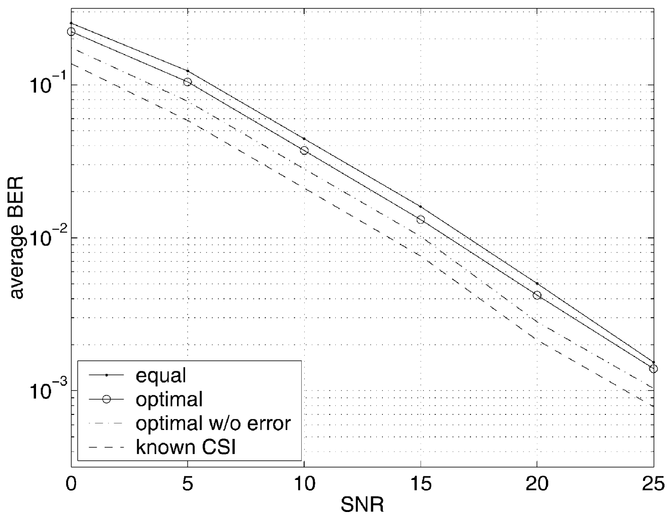

Fig. 5. Average BER versus SNR (i.i.d. channel taps).

channel capacity and capacity bounds, which highlights the validity of our lower bound on average channel capacity.

VI. Conclusion

We considered wireless OFDM-like systems with decoupled information-bearing from pilot symbols transmitted over FIR Rayleigh channels having independent taps with arbitrary power profiles. We proved that placing pilots in an equispaced and equipowered manner minimizes the mean-square channel estimation error. From the capacity lower bound of our system with equispaced and equipowered pilots, we determined the number of the pilots and, for a given transmit power budget, we also derived the optimal percentages of power to be distributed between information and pilot symbols. We found that in non-i.i.d. Rayleigh-fading channels, channel estimation with optimal power allocation incurs asymptotically (high SNR) at most 3 dB loss in channel capacity; while in i.i.d. Rayleigh-fading channels, the loss is at most 6 dB at high SNR.

Appendix I
Proof of (21)

We will first prove the following lemma, from which (21) follows immediately.

Lemma 1: For an $N \times N$ positive-definite matrix $\boldsymbol{A}$ with (m,n)th entry $a_{m,n}$, it holds that

$$\mathrm{trace}(\boldsymbol{A}^{-1}) \geq \sum_{n=1}^{N} \frac{1}{a_{n,n}} \tag{40}$$

where the equality is attained if and only if $\boldsymbol{A}$ is diagonal.

Proof: We partition $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{a} \\ \boldsymbol{a}^{\mathcal{H}} & a_{N,N} \end{bmatrix} \tag{41}$$

where $\boldsymbol{A}_{11}$ is an $(N-1)\times(N-1)$ matrix and $\boldsymbol{a}$ is an $(N-1)\times 1$ vector. Using matrix inversion of the partitioned matrix, we have

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{a}\boldsymbol{a}^{\mathcal{H}}\boldsymbol{A}_{11}^{-1}/d & -\boldsymbol{A}_{11}^{-1}\boldsymbol{a}/d \\ -(\boldsymbol{A}_{11}^{-1}\boldsymbol{a}/d)^{\mathcal{H}} & 1/d \end{bmatrix} \tag{42}$$

where $d = a_{N,N} - \boldsymbol{a}^{\mathcal{H}}\boldsymbol{A}_{11}^{-1}\boldsymbol{a}$. Because $\boldsymbol{A} > 0$ and $d > 0$, we find that $1/d \geq 1/a_{N,N}$. We also have $\boldsymbol{A}_{11}^{-1}\boldsymbol{a}\boldsymbol{a}^{\mathcal{H}}\boldsymbol{A}_{11}^{-1}/d \geq 0$. It follows that

$$\begin{aligned}\mathrm{trace}\,\boldsymbol{A}^{-1} &= \mathrm{trace}\left(\boldsymbol{A}_{11}^{-1} + \frac{1}{d}\boldsymbol{A}_{11}^{-1}\boldsymbol{a}\boldsymbol{a}^{\mathcal{H}}\boldsymbol{A}_{11}^{-1}\right) + \frac{1}{d} \\ &\geq \mathrm{trace}\,\boldsymbol{A}_{11}^{-1} + \frac{1}{a_{N,N}}\end{aligned}$$

where the equality holds if and only if $\boldsymbol{a} = \boldsymbol{0}$. Repeating the argument above, we obtain (40). □

Appendix II
Proof of $E\{|\Delta H(W^i)|^2\} = \sigma_{\Delta h}^2$

Since $(\boldsymbol{R}_h^{-1} + \boldsymbol{B}^{\mathcal{H}}\boldsymbol{B}/\sigma_w^2)^{-1}$ is diagonal under C3 (cf. the arguments following (21)), we have

$$\begin{aligned}&E\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\mathcal{H}}\} \\ &\quad = \mathrm{diag}\left(\frac{\sigma_{h_0}^2\sigma_w^2}{\sigma_w^2 + \mathcal{P}_b\sigma_{h_0}^2}, \frac{\sigma_{h_1}^2\sigma_w^2}{\sigma_w^2 + \mathcal{P}_b\sigma_{h_1}^2}, \ldots, \frac{\sigma_{h_L}^2\sigma_w^2}{\sigma_w^2 + \mathcal{P}_b\sigma_{h_L}^2}\right).\end{aligned}$$

It follows from

$$[\Delta H(W^{i_0}), \Delta H(W^{i_1}), \ldots, \Delta H(W^{i_{M-1}})]^T = N^{\frac{1}{2}}\breve{\boldsymbol{F}}_{\bar{L}}\Delta\boldsymbol{h}$$

and (17) that for $i_n \in \mathcal{I}$

$$\begin{aligned}E\{|\Delta H(W^{i_n})|^2\} &= N\boldsymbol{f}_{\bar{L},n}^{\mathcal{H}}E\{\Delta\boldsymbol{h}\Delta\boldsymbol{h}^{\mathcal{H}}\}\boldsymbol{f}_{\bar{L},n} \\ &= \sum_{l=0}^{L}\frac{\sigma_{h_l}^2\sigma_w^2}{\sigma_w^2 + \mathcal{P}_b\sigma_{h_l}^2} = \sigma_{\Delta h}^2\end{aligned}$$

where $\boldsymbol{f}_{\bar{L},n}^{\mathcal{H}}$ is a row vector equal to the i_nth row of $\boldsymbol{F}_{\bar{L}}$. This completes the proof. □

Appendix III
Derivation of (36)

We substitute

$$\sigma_{\Delta h}^2 = \bar{L}\sigma_h^2\sigma_w^2/(\sigma_w^2 + \mathcal{P}_b\sigma_h^2)$$

and

$$\sigma_H^2 - \sigma_{\Delta h}^2 = \bar{L}\mathcal{P}_b\sigma_h^4/(\sigma_w^2 + \mathcal{P}_b\sigma_h^2)$$

into (27) to obtain

$$\begin{aligned}\rho &= \frac{\bar{L}\mathcal{P}_b\sigma_h^4/(\sigma_w^2 + \mathcal{P}_b\sigma_h^2)}{\bar{L}\sigma_h^2\sigma_w^2/(\sigma_w^2 + \mathcal{P}_b\sigma_h^2) + M\sigma_w^2/\mathcal{P}_s} \\ &= \frac{\bar{L}\sigma_h^2}{\sigma_w^2}\frac{\mathcal{P}_b\mathcal{P}_s}{\bar{L}\mathcal{P}_s + M(\sigma_w^2/\sigma_h^2 + \mathcal{P}_b)}.\end{aligned}$$

From definitions $\mathcal{P}_s = \alpha\mathcal{P}$, $\mathcal{P}_b = (1-\alpha)\mathcal{P}$ and $\rho_{\mathrm{snr}} = \bar{L}\sigma_h^2\mathcal{P}/(M\sigma_w^2)$, we arrive at

$$\rho = \rho_{\mathrm{snr}}\frac{\alpha(1-\alpha)}{\bar{L}\alpha + M[\bar{L}/(M\rho_{\mathrm{snr}}) + 1 - \alpha]} \tag{43}$$

which is clearly equivalent to (36). □

Acknowledgment

The authors wish to thank the reviewers for their valuable feedback.

References

[1] S. Adireddy and L. Tong, "Detection with embedded known symbols: Optimal symbol placement and equalization," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 5, Istanbul, Turkey, June 2000, pp. 2541–2543.

[2] ——, “Optimal embedding of known symbols for OFDM,” in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001.
[3] S. Adireddy, L. Tong, and H. Viswanathan, “Optimal placement of training for frequency-selective block-fading channels,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 2338–2353, Aug. 2002.
[4] N. Al-Dhahir and J. M. Cioffi, “Block transmission over dispersive channels: Transmit filter optimization and realization, and MMSE-DFE receiver performance,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 137–160, Jan. 1996.
[5] E. Biglieri, J. Proakis, and S. Shamai (Shitz), “Fading channels: Information-theoretic and communications aspects,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2619–2692, Oct. 1998.
[6] 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.
[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[8] ETSI, “Broadband Radio Access Networks (BRAN); HIPERLAN Type 2 Technical Specification Part 1—Physical Layer,” report, DTS/BRAN030003-1, 1999.
[9] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
[10] B. Hassibi and B. Hochwald, “How much training is needed in multiple-antenna wireless links?,” *IEEE Trans. Inform. Theory*. Available: [Online] at http://mars.bell-labs.com/cm/ms/what/mars/index.html, to be published.
[11] P. Ho and J. H. Kim, “Pilot symbol-assisted detection of CPM schemes operating in fast fading channels,” *IEEE Trans. Commun.*, vol. 44, pp. 337–347, Mar. 1996.
[12] P. Höher and F. Tufvesson, “Channel estimation with super imposed pilot sequence,” in *Proc. GLOBECOM Conf.*, Rio de Janeiro, Brazil, Dec. 1999, pp. 2162–2166.
[13] G. K. Kaleh, “Channel equalization for block transmission systems,” *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1728–1736, Jan. 1995.
[14] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993, vol. 1.
[15] T. Keller and L. Hanzo, “Adaptive multicarrier modulation: A convenient framework for time-frequency processing in wireless communications,” *IEEE Commun. Mag.*, vol. 88, pp. 611–640, May 2000.
[16] A. Lapidoth and S. Shamai (Shitz), “Fading channels: How perfect need perfect side information be?,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 1118–1134, May 2002.
[17] T. L. Marzetta and B. M. Hochwald, “Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading,” *IEEE Trans. Inform. Theory*, pp. 139–157, Jan. 1999.
[18] M. Médard, “The effect upon channel capacity in wireless communication of perfect and imperfect knowledge of the channel,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 933–946, May 2000.
[19] R. Negi and J. Cioffi, “Pilot tone selection for channel estimation in a mobile OFDM system,” *IEEE Trans. Consumer Electron.*, vol. 44, pp. 1122–1128, Aug. 1998.
[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] A. Sabharwal, E. Erkip, and B. Aazhang, “On channel state information in multiple antenna block fading channels,” in *Proc. 2000 Int. Symp. Information Theory and its Applications*, Honolulu, HI, Nov. 2000, pp. 116–119.
[22] D. Samardzija and N. Mandayam, “Pilot assisted estimation of MIMO fading channel response and achievable data rates,” in *Proc. DIMACS Workshop on Signal Processing for Wireless Transmission*, Piscataway, NJ, Oct. 7–9, 2002.
[23] H. Sari, G. Karam, and I. Jeanclaude, “Transmission techniques for digital terrestrial TV broadcasting,” *IEEE Commun. Mag.*, vol. 33, pp. 100–109, Feb. 1995.
[24] A. Scaglione, S. Barbarossa, and G. B. Giannakis, “Filterbank transceivers optimizing information rate in block transmissions over dispersive channels,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 1019–1032, Apr. 1999.
[25] A. Scaglione, G. B. Giannakis, and S. Barbarossa, “Redundant filterbank precoders and equalizers parts I & II,” *IEEE Trans. Signal Processing*, vol. 47, pp. 1988–2022, July 1999.
[26] İ. E. Telatar, “Capacity of multiple-antenna Gaussian channels,” *Europ. Trans. Telecommun.*, vol. 10, pp. 585–595, Nov./Dec. 1999.
[27] Z. Wang and G. B. Giannakis, “Wireless multicarrier communications: Where Fourier meets Shannon,” *IEEE Signal Processing Mag.*, vol. 47, no. 3, pp. 29–48, May 2000.
[28] ——, “Complex-field coding for OFDM over fading wireless channels,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 707–720, Mar. 2003.
[29] H. Weingarten and Y. Steinberg, “Gaussian codes and the scaled nearest neighbor decoder in fading multi-antenna channels,” in *Proc. 39th Allerton Conf.*, Monticello, IL, Oct. 3–5, 2001, pp. 825–834.
[30] L. Zheng and D. N. C. Tse, “Information theoretic limits for non-coherent multi-antenna communications,” in *Proc. IEEE Wireless Communications and Networking Conf.*, Sept. 2000, pp. 18–22.

Rate of Convergence to Poisson Law in Terms of Information Divergence

Peter Harremoës, *Member, IEEE,* and Pavel S. Ruzankin

***Abstract*—The precise bounds on the information divergence from a binomial distribution to the accompanying Poisson law are obtained. As a corollary, an upper bound for the total variation distance between the distributions is established, which in a large range of change of the parameters of the distributions is better than those already known.**

***Index Terms*—Binomial distribution, information divergence, Kullback–Leibler divergence, Poisson approximation, Stirling numbers.**

I. Introduction and the Main Results

Let Q be the binomial distribution with parameters n, p and P be the Poisson law with parameter np. The information divergence (or the Kullback–Leibler divergence or the relative entropy) from the distribution Q to the distribution P is

$$D(Q\|P) = \sum_{k=0}^{n} Q(\{k\}) \log \frac{Q(\{k\})}{P(\{k\})}$$

and the total variation distance between Q and P is

$$V(Q,P) = \sum_{k=0}^{n} |Q(\{k\}) - P(\{k\})|.$$

These two quantities satisfy the Pinsker inequality (see [13] and [4])

$$\frac{1}{2}(V(Q,P))^2 \leq D(Q\|P). \tag{1}$$

For the means of the laws being fixed and equal, the information divergence from the binomial distribution to the Poisson distribution is small [6]. Consider a statistical test for the null-hypothesis that a sample

Manuscript received March 9, 2003; revised March 16, 2004. The work of P. Harremoës was supported by a postdoctoral fellowship from the Villum Kann Rasmussen Foundation, by INTAS (Project 00-738), and by the Danish Natural Science Research Council. The work of P. S. Ruzankin was supported in part by the Russian President's Grants MK-2600.2003.01 and NSh-2139.2003.01, the RFBR Projects 02-01-00902 and 03-01-00459, and INTAS (Project 03-51-5018).

P. Harremoës is with the Department of Mathematics, University of Copenhagen, DK-2100 Copenhagen, Denmark (e-mail: moes@math.ku.dk).

P. S. Ruzankin is with the Sobolev Institute of Mathmatics of the Russian Academy of Sciences, Novosibirsk, 630090 Russia (e-mail: ruzankin@math.nsc.ru).

Communicated by R. W. Yeung, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2004.833364