ECE 361: Digital Communications

Lecture 16: Transmitter-Centric ISI Harnessing: Orthogonal Frequency Division Modulation (OFDM)

Introduction

In the previous lecture, we took a first order transmitter-centric approach to dealing with ISI: the focus was on eliminating the effects of ISI. In this lecture we take a more balanced view: harnessing the benefits of ISI instead of just treating it as interference, continuing our transmitter-centric approach. Our goal is to get the full benefit of the fact that multiple delayed copies of the transmit symbols appear at the receiver while still employing a receiver no more complicated than the one used over an AWGN channel. While this seems to be a tall order, we will see a remarkable scheme that achieves exactly this. In a sense, it is a natural culmination of the various ISI mitigation techniques we have seen in the course of the past several lectures. The scheme that converts the frequency selective ISI channel into a plain AWGN channel is known as *orthogonal frequency division modulation* (OFDM) and is the main focus of this lecture.

An Ideal Situation

Consider the frequency selective model that we have been working with as a good approximation of the wireline channel:

$$y[m] = \sum_{\ell=0}^{L-1} h_\ell x[m-\ell] + w[m], \quad m \geq 1. \tag{1}$$

Since we know how to communicate reliably over an AWGN channel (cf. Lecture 7), it would be ideal (not to mention, easy) if the channel with ISI can *somehow* (by appropriate transmitter and reciever operations) be converted into an AWGN one: say,

$$y[m] = \hat{h}x[m] + w[m], \quad m \geq 1. \tag{2}$$

In such a case, we could simply and readily use the transmitter and receiver techniques developed already for the AWGN channel (available "off-the-shelf", so to say).

While this is asking for a bit too much, we will see that we can get somewhat close: indeed, we will convert the ISI channel in Equation (1) into a *collection* of AWGN channels, each of different noise energy level:

$$\hat{y}[N_c k + n] = \hat{h}_n \hat{x}[N_c k + n] + \hat{w}[N_c k + n], \quad k \geq 0, \quad n = 0 \ldots N_c - 1. \tag{3}$$

The idea is that the time index m is replaced by $N_c k + n$. The inputs are voltages $\hat{x}$. The additive noise $\hat{w}[\cdot]$ is white Gaussian (zero mean and variance σ^2). We observe that there are N_c different AWGN channels, one for each $n = 0, \ldots, N_\mathrm{c} - 1$. We can make two further observations:

- each of the N_c AWGN channels has a *different* operating SNR: the n^{th} channel has an SNR equal to $\hat{h}_n^2 \mathsf{SNR}$ where SNR is, as usual, the ratio of the transmit energy to the noise energy;
- each of the N_c AWGN channels is available for use only a *fraction* $\frac{1}{N_\mathrm{c}}$ of the time.

Such a collection of non-interfering AWGN channels is called a *parallel* AWGN channel. The individual AWGN channels within the collection are known as *sub-channels*. Our understanding of efficient reliable communication over the AWGN channel suggests a natural strategy to communicate over the parallel AWGN channel as well: we can split the information bits so that we communicate over each sub-channel *separately*. The only choice remaining is how to split the total power budget amongst the sub-carriers, say power P_n to the n^{th} sub-carrier, so that

$$\sum_{n=0}^{N_\mathrm{c}-1} P_n = P, \tag{4}$$

where P is the total power budget for reliable communication. With a transmit power constraint of P_n, the overall SNR of the n^{th} sub-channel is

$$\frac{P_n \hat{h}_n^2}{\sigma^2}. \tag{5}$$

Thus, with an appropriate coding and decoding mechanism reliable communication is possible (cf. Lecture 7) on the n^{th} AWGN sub-channel at rate

$$R_n = \frac{1}{2N_\mathrm{c}} \log_2 \left(1 + \frac{P_n \hat{h}_n^2}{\sigma^2} \right), \tag{6}$$

measured in bits/symbol. The factor of $1/N_\mathrm{c}$ in the rate appears because each of the sub-channels is available for use only a fraction $1/N_\mathrm{c}$ of the time. The total rate of reliable communication is

$$\sum_{n=0}^{N_\mathrm{c}-1} R_n = \frac{1}{2N_\mathrm{c}} \sum_{n=0}^{N_\mathrm{c}-1} \log_2 \left(1 + \frac{P_n \hat{h}_n^2}{\sigma^2} \right). \tag{7}$$

We can now split the power to *maximize* the rate of reliable communication over the parallel AWGN channel:

$$\max_{P_n \geq 0, \sum_{n=0}^{N_\mathrm{c}-1} P_n = P} \frac{1}{2N_\mathrm{c}} \sum_{n=0}^{N_\mathrm{c}-1} \log_2 \left(1 + \frac{P_n \hat{h}_n^2}{\sigma^2} \right). \tag{8}$$

The optimal power split can be derived explicitly and is explored in a homework exercise. The main property of this optimal power split is that

> the larger the "quality" of a sub-channel, the more the power that is allocated to it, and hence the larger the corresponding data rate of reliable communication.

In the rest of this lecture we will see a "high level view" of how to go from the ISI channel (cf. Equation (1) to the parallel AWGN channel (cf. Equation (3)).

Convolution in Matrix Notation

Let's consider communicating over a block of N_c time instants over a L-tap wireline channel. Recall that the received voltage at the output of the wireline channel at time m is

$$y[m] = \sum_{l=0}^{L-1} h_l x[m-l] + w[m] \qquad m \geq 1. \tag{9}$$

Here, $x[m]$ and $w[m]$ are the transmitted and the noise voltages at time instant m, respectively. The wireline channel without the additive noise is a simple linear (finite impulse response) filter. It *convolves* the input voltage vector with the channel vector. Adding the noise vector gives the received voltage vector. Since convolution is, at its essence, a *linear* transformation, we may employ the matrix notation to depict the input-output relation over a wireline channel:

$$\begin{bmatrix} y[1] \\ y[2] \\ y[3] \\ \vdots \\ \vdots \\ y[N_c] \end{bmatrix} = \begin{bmatrix} h_0 & 0 & 0 & 0 & \cdots & 0 \\ h_1 & h_0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & h_{L-1} & h_{L-2} & \cdot & h_0 \end{bmatrix} \begin{bmatrix} x[1] \\ x[2] \\ \vdots \\ x[N_c] \end{bmatrix} + \begin{bmatrix} w[1] \\ w[2] \\ w[3] \\ \vdots \\ \vdots \\ w[N_c] \end{bmatrix}. \tag{10}$$

We can write Equation 10 in short hand notation as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{11}$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{w}$ are the input, output and noise vectors (all of dimension $N_c \times 1$), respectively. The index k here refers to different N_c dimensional blocks of input voltage vectors $\mathbf{x}$ that could be sent over the wireline channel. The channel matrix $\mathbf{H}$ is a $N_c \times N_c$ dimensional matrix. The matrix $\mathbf{H}$ has an important feature: the diagonal elements are all identical. Such a matrix is said to be *Toeplitz*.

Communicating over a Diagonal Matrix

The channel expressed in Equation (11) is a *vector* version of the AWGN channel we are very familiar with. If the linear transformation $\mathbf{H}$ was diagonal, then we would have a *parallel* AWGN channel. Communicating over such a channel is relatively straightforward: we can code over each "sub-channel" reliably using coding and decoding methods developed for AWGN channels; this has been explained in detail in the previous lecture. However, the $\mathbf{H}$ that results from the convolution operation is diagonal exactly when there is just one tap (i.e., $L = 1$). So, we cannot really expect $\mathbf{H}$ to be diagonal at all. Since we know how to deal with a diagonal matrix, a question that naturally arises is:

> Can we perform some linear transformation on $\mathbf{H}$ so that the resulting matrix is diagonal?

If so, then the linear transformations on $\mathbf{H}$ can be absorbed into linear processing at the transmitter and the receiver. Concretely, let $\mathbf{Q}_1$ denote the transformation done at the transmitter: the "data" vector $\tilde{\mathbf{x}}$ and the transmit voltage vector $\mathbf{x}$ are related by

$$\mathbf{x} = \mathbf{Q}_1\tilde{\mathbf{x}}. \tag{12}$$

Analogously, let matrix $\mathbf{Q}_2$ denote the transformation on the received voltage vector $\mathbf{y}$:

$$\tilde{\mathbf{y}} = \mathbf{Q}_2\mathbf{y} \tag{13}$$

We can then write the *effective* channel between the data vector $\tilde{\mathbf{x}}$ and vector $\tilde{\mathbf{y}}$ as:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\tilde{\mathbf{x}} + \tilde{\mathbf{w}} \tag{14}$$

where

$$\begin{aligned} \tilde{\mathbf{H}} &= \mathbf{Q}_2\mathbf{H}\mathbf{Q}_1 && (15)\\ \tilde{\mathbf{w}} &= \mathbf{Q}_2\mathbf{w}. && (16) \end{aligned}$$

We want to choose $\mathbf{Q}_1$ and $\mathbf{Q}_2$ such that the matrix $\tilde{\mathbf{H}}$, the effective channel between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$, is diagonal. But we would like to do this while still not changing the statistics of the vector noise $\tilde{\mathbf{w}}$ (i.e., the entries of $\tilde{\mathbf{w}}$ are also i.i.d. Gaussian random variables). This way, we would have arrived at a parallel AWGN channel. To summarize: our goal is to find the matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ such that

1. $\mathbf{Q}_2\mathbf{H}\mathbf{Q}_1$ is diagonal;

2. $\tilde{\mathbf{w}}$ is an i.i.d. Gaussian random vector.

Linear Transformation of a Matrix Into a Diagonal Form

Linear algebra is a mathematical field of study that answers questions of the type above. It would be a big detour for us to wander along this path, but here is the classical text book that would satiate all but the most curious graduate student:

> R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

The second condition, that $\mathbf{w}$ and $\tilde{\mathbf{w}}$ both be i.i.d. Gaussian is a relatively mild one. In fact, this is readily satisfied by all matrices $\mathbf{Q}_2$ with the property that

$$\mathbf{Q}_2^T \mathbf{Q}_2 = \mathbf{I}. \tag{17}$$

Here we have denoted the identity matrix by $\mathbf{I}$. A simple calculation verifying this claim for $N_c = 2$ is available in the appendix. The first condition, requiring $\tilde{\mathbf{H}}$ to be diagonal is more involved. In fact, the most fundamental result in all of linear algebra, the so-called *singular value decomposition* (SVD) theorem, answers this question in the affirmative: all matrices $\mathbf{H}$ can be linearly transformed into diagonal matrices $\tilde{\mathbf{H}}$. The choice of $\mathbf{Q}_1, \mathbf{Q}_2$ that achieve this transformation, in general, depend closely on the original matrix $\mathbf{H}$. However, we are dealing with a Toeplitz matrix $\mathbf{H}$ and there is some chance to have specific structure in the appropriate choice of $\mathbf{Q}_1$ and $\mathbf{Q}_2$. Indeed, there is strong structure when N_c is large. This aspect is explored in the theory of Toeplitz forms, a topic well beyond the boundaries of mathematics we are using. Then again, it is hard to say when one of you might need this knowledge. So, here is the classical reference on the topic:

> U. Grenander, *Toeplitz Forms and Their Applications*, Chelsea Publication Company, 2nd edition, 1984.

The structure in $\mathbf{Q}_1, \mathbf{Q}_2$ for large enough N_c is a bit hard to describe. With a bit of modification, however, there is a very clean solution. This is the OFDM solution, the focus of the next lecture.

Looking ahead

We took a fundamental block-communication view of the ISI channel. We saw that it is possible to convert the channel into a *parallel* AWGN channel. Reliable communication over the parallel AWGN channel is simple: we send separate packets over the separate AWGN sub-channels. We have the flexibility of allocating powers (and appropriate data rates) in our communication over the sub-channels. A good strategy would be to allocate powers in proportion to the sub-channel quality. In the next lecture we will see a concrete and simple algorithmic method to convert the ISI channel into a parallel AWGN channel.

Appendix

Set $N_c = 2$. We are interested in properties of the 2×2 matrix $\mathbf{Q}_2$ such that both $\mathbf{w}$ and $\tilde{\mathbf{w}}$ are i.i.d. Gaussian:

$$\begin{aligned}
\begin{bmatrix} \tilde{w}_1 \\ \tilde{w}_2 \end{bmatrix} &= \mathbf{Q}_2 \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} && (18) \\
\begin{bmatrix} \tilde{w}_1 \\ \tilde{w}_2 \end{bmatrix} &= \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}. && (19)
\end{aligned}$$

We know that w_1 and w_2 are i.i.d. Gaussian random variables with mean 0 and variance σ^2. We want $\tilde{w}_1$ and $\tilde{w}_2$ to also be i.i.d., i.e., they should necessarily have same variance and should be uncorrelated. (Note that this is also the sufficient condition as $\tilde{w}_i$ are Gaussian.)

$$\begin{aligned}
\mathrm{Var}(\tilde{w}_1) &= (a_1^2 + b_1^2)\sigma^2 && (20) \\
\mathrm{Var}(\tilde{w}_2) &= (a_2^2 + b_2^2)\sigma^2 && (21) \\
\Rightarrow (a_1^2 + b_1^2) &= (a_2^2 + b_2^2) = 1 \quad \text{(say)}. && (22)
\end{aligned}$$

The correlation $\mathbb{E}(\tilde{w}_1 \tilde{w}_2)$ is

$$\begin{aligned}
\mathbb{E}(\tilde{w}_1 \tilde{w}_2) &= \mathbb{E}[(a_1 w_1 + b_1 w_2)(a_2 w_1 + b_2 w_2)] && (23) \\
&= (a_1 a_2 + b_1 b_2)\sigma^2. && (24)
\end{aligned}$$

For $\tilde{w}_1$ and $\tilde{w}_2$ to be uncorrelated, we want

$$a_1 a_2 + b_1 b_2 = 0 \tag{25}$$

From Equation (22), we can choose the vectors $[a_1, b_1]$ and $[a_2, b_2]$ as

$$\mathbf{Q}_2 = \begin{bmatrix} \cos\theta & \sin\theta \\ \cos\phi & \sin\phi \end{bmatrix}. \tag{26}$$

From Equation 25, we want

$$\begin{aligned}
\cos\theta \cos\phi + \sin\theta \sin\phi &= 0 && (27) \\
\cos(\theta - \phi) &= 0 && (28) \\
\theta = \phi + \frac{\pi}{2} & && (29)
\end{aligned}$$

Hence

$$\mathbf{Q}_2 = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \tag{30}$$

which has the property that $\mathbf{Q}_2^T \mathbf{Q}_2 = \mathbf{I}$ for all values of θ, as can be directly verified.