ECE461: Digital Communications

Lecture 5: Energy Efficient Communication

Introduction

So far, we have seen that block communication (using the simple repetition coding) can improve the reliability of communication, significantly over and above that possible with sequential communication. This is particularly true when we want to communicate a large amount of data. But this has come at a high cost: specifically we can get arbitrarily reliable communication, but

- the data rate (number of bits per unit time sample) goes to zero. Specifically, we know from Lecture 4 that the data rate is

$$\frac{B}{n} = o\left(\frac{\log_2 n}{2n}\right), \tag{1}$$

  where we used the notation $o(f(n))$ to denote a function of n that has the property that

$$\lim_{n\to\infty} o(f(n)) = 0. \tag{2}$$

  Simply put we can think of the data rate of reliable communication with repetition coding as approximately

$$\frac{\log_2 n}{2n} \tag{3}$$

  which is very small for large n. For a large data packet (of size, say B), we need an amount of time approximately 2^{2B} to communicate it reliably using repetition coding!

- it is very energy inefficient. Here, we have defined the energy efficiency is defined as the amount of energy (in Joules) consumed per bit that is reliably communicated. In the repetition coding scheme, using n time instants we are using a total energy nE. Further, we need 2^{2B} time samples to send B bits reliably. So we use up energy proportional to 2^{2B} and thus the energy efficiency is

$$\frac{2^{2B}E}{B}. \tag{4}$$

  For large data size B, this goes to infinity: the repetition coding scheme is hardly energy efficient.

In summary,

> repetition coding significantly improved the reliability of communication over sequential communication, particularly for large data packets, but at the cost of zero data rate and zero energy efficiency.

This is in stark contrast to sequential communication that had *non-zero* data rate and energy efficiency: after all, we keep transmitting new information bits at every time sample (so the data rate is non-zero) and we only use a finite energy at any time sample (so the energy efficiency is also non-zero).

Question: Can we have the desirable features of sequential communication, non-zero data rate and energy efficiency, while ensuring that the reliability is very good? In other words, is there a free breakfast, lunch and dinner?

Well, the short answer is *yes*. The long answer is that the block communication scheme that does it is quite involved. It *actually* is rocket-science (almost!). We will spend several lectures trying to figure out what it takes to reliably communicate at non-zero data rates and non-zero energy efficiency. It is a remarkable success story that has drawn various aspects of electrical engineering: role of modeling, asking the right questions, mathematical abstraction, mathematical legerdemain, algorithmic advances and finally advances in circuits to handle the computational complexity of the algorithms involved.

We take a small step towards this story in this lecture by focusing on reliably communicating at non-zero energy efficiency (while not worrying about non-zero data rate).

Peak Energy vs Average Energy Constraints

Before we introduce our energy-efficient communication scheme, it helps to consider a slightly different type of energy constraint. So far, our constraint has been that our voltage level at any time sample is bounded in magnitude (i.e., it has to be within a limit of $\pm\sqrt{E}$ Volts in our previous notation). This comes from a constraint on the instantaneous energy at every time sample. In several communication scenarios, the constraint is on the *average* energy, averaged over many time instants. This is the constraint whenever an electronic device is rated in Watts (unit of power) measured in Joules/second. For instance, we could restrict the total power transmitted to P and this would mean that the sum of the square of the voltages transmitted in n time samples is no more than nP. In this lecture, we will focus on reliable block communication when there is a power constraint (or an average energy constraint), as opposed to instantaneous (or peak) energy constraint.

Suppose we want to transmit B bits reliably and energy efficiently. This means that we want to use a finite amount of energy per bit transmitted: so the total energy allowed to transmit bits is directly proportional to B. Let us denote the energy allowed per bit to be $\mathcal{E}_b$ (a finite value), so the total energy allowed is $B\mathcal{E}_b$.

Since the data rate is not the focus here, let us consider using a lot of time samples to transmit the bits (just as we needed to do in the repetition coding scheme). Specifically, suppose we use 2^B time instants (still an improvement over the repetition coding scheme which required 2^{2B} time instants). The data rate is surely very small for large B, but as we said, let us not worry about that now. Let us number the time samples from 0 through $2^B - 1$. Now every data packet with B bits can be made to correspond *exactly* to an integer between 0 and $2^B - 1$: one way to do this is think of the data packet (string of bits) as the binary expansion of an integer. That integer is surely somewhere between 0 and $2^B - 1$.

Now we see the possible reason for choosing the number of time samples *equal* to the number of different possible realizations (2^B) of B bits. It is to allow for the following *position* modulation scheme: look at the information packet of B bits and convert it to

an integer between 0 and $2^B - 1$, say denoted by k. Now we transmit nothing (or zero in mathematical terms) at all *except* the k^{th} time samples when we use up all the energy we have: we do this by transmitting the voltage $+\sqrt{B\mathcal{E}_b}$.

This type of modulation is very different from the types of modulation schemes seen in earlier lectures: here the information is in the *position* of a large voltage rather than the specific *amplitude* of a voltage. As such, this type of communication is referred to as position modulation and is in contrast to the earlier amplitude modulation schemes. In terms of the transmit voltage vector (of length 2^B), it looks like a vector

$$\left(0, 0, \ldots, \sqrt{B\mathcal{E}_b}, 0, \ldots, 0\right) \tag{5}$$

where the only non-zero entry $B\mathcal{E}_b$ is at the k^{th} location in the vector. Figure illustrates the position modulation scheme for $B = 2$.

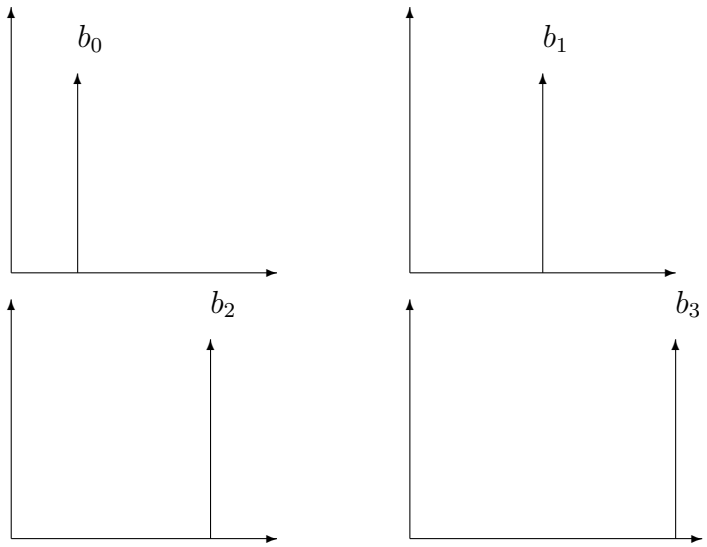

Figure 1: Position modulation: horizontal axis represents time samples and the vertical axis the voltages transmitted.

ML Rule

How do we expect the receiver to decide on which position the voltage might have been sent? Clearly, taking the average of the voltages (like we did earlier for repetition coding) is not going to help. A natural idea is that since Gaussian noise is more likely to be small (near its mean of zero) than large, we just pick that time when the received voltage is the largest. This very intuitive rule is indeed what the ML rule also is. We work this out below, for completeness and as a formal verification of our engineering intuition.

The receiver receives 2^B voltages

$$y[m] = x[m] + w[m], \quad m = 1, 2, \ldots, 2^B. \tag{6}$$

The ML rule involves calculating the likelihood of the voltages received for each possible position of where the non-zero voltage was transmitted. Suppose $k = 1$ is the index where the non-zero voltage is transmitted, i.e., the transmit voltage vector is

$$x[m] = \begin{cases} \sqrt{B\mathcal{E}_b} & m = 1 \\ 0 & \text{else.} \end{cases} \tag{7}$$

The likelihood of receiving the voltages $a_1, \ldots a_{2^B}$ is

$$\begin{aligned} L_1 &= f_{w_1}\left(a_1 - \sqrt{B\mathcal{E}_b}\right) f_{w_2}(a_2) \ldots f_{w_{2^B}}(a_{2^B}) && (8) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2^B} \exp\left(-\frac{\left(a_1 - \sqrt{B\mathcal{E}_b}\right)^2 + \sum_{m=2}^{2^B} a_m^2}{2\sigma^2}\right) && (9) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2^B} \exp\left(-\frac{\sum_{m=1}^{2^B} a_m^2}{2\sigma^2} - \frac{B\mathcal{E}_b}{2\sigma^2}\right) \exp\left(\frac{a_1\sqrt{B\mathcal{E}_b}}{\sigma^2}\right). && (10) \end{aligned}$$

By the symmetry of the modulation scheme in the transmit index k, we have (following Equation 10)

$$L_k = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2^B} \exp\left(-\frac{\sum_{m=1}^{2^B} a_m^2}{2\sigma^2} - \frac{B\mathcal{E}_b}{2\sigma^2}\right) \exp\left(\frac{a_k\sqrt{B\mathcal{E}_b}}{\sigma^2}\right). \tag{11}$$

The ML rule picks that index k which has the largest likelihood L_k. The first two terms in the formula (cf. Equation 11) for L_k are *independent* of the index k. So, we can focus on just the third term. There we see that maximizing it is simply a matter of picking k such that a_k is the largest.

Reliability with Position Modulation

Suppose, again, that the index k where the non-zero voltage was transmitted is 1. Since there is complete symmetry of the modulation scheme with respect to k, we can just calculate the probability of error for this value of k and it will be the same unreliability level for all other choices. Now, the ML rule makes a mistake whenever at least one of the received voltages $a_2, \ldots, a_{2^B}$ is larger than a_1. Denoting the event

$$\text{Error Event}_{1j} = \{a_j > a_1\}, \quad j = 2 \ldots 2^B \tag{12}$$

we see that the error event when $k = 1$ is their *union*:

$$\text{Error} = \bigcup_{j=2}^{2^B} \text{Error Event}_{1j}. \tag{13}$$

It turns out that the probability of the error event is somewhat complicated to calculate directly. We can find an upper bound to it easily enough though, which itself will be easier to evaluate.

The probability of making an error is now upper bounded by the sum of the probabilities of the *pair-wise* events: indeed,

$$\mathbb{P}\left[\text{Error}\right] \leq \sum_{j=2}^{2^B} \mathbb{P}\left[\text{Error Event}_{1j}\right]. \tag{14}$$

The idea is that if we can bound the right-hand side of this equation by a small enough number, then the unreliability level of communication with position modulation itself is less than that small enough number. Such a way of bounding the error probability is known as the *union bound*.

How do we calculate the pair-wise error event probability $\mathbb{P}\left[\text{Error Event}_{1j}\right]$? We only need to focus on what happens at the time samples 1 and j. The received voltages at these two time samples look as follows:

$$\begin{aligned} y[1] &= \sqrt{B\mathcal{E}_b} + w[1], & (15)\\ y[j] &= w[j]. & (16) \end{aligned}$$

Now the pair-wise error probability is

$$\begin{aligned} \mathbb{P}\left[\text{Error Event}_{1j}\right] &= \mathbb{P}\left[y[j] > y[1]\right] & (17)\\ &= \mathbb{P}\left[w[j] - w[1] > \sqrt{B\mathcal{E}_b}\right]. & (18) \end{aligned}$$

Observe that the difference of two independent Gaussian noises (with the same mean and variance) also has Gaussian statistics but has *twice* the variance as the original ones. So, the difference $w[j] - w[1]$ is Gaussian with zero mean and variance $2\sigma^2$. Now we have a simple expression for the pair-wise error probability: continuing from Equation 18

$$\mathbb{P}\left[\text{Error Event}_{1j}\right] = Q\left(\frac{\sqrt{B\mathcal{E}_b}}{\sqrt{2\sigma^2}}\right), \quad j = 2 \ldots 2^B. \tag{19}$$

We can substitute Equation 19 in Equation 14 to arrive at an upper bound to the unreliability level of communication with position modulation:

$$\mathbb{P}\left[\text{Error}\right] \leq \left(2^B - 1\right) Q\left(\frac{\sqrt{B\mathcal{E}_b}}{\sqrt{2\sigma^2}}\right). \tag{20}$$

Using the usual upper bound to the $Q(\cdot)$ function (cf. Homework 1), we can get a further upper bound to the error probability:

$$\begin{aligned} \mathbb{P}\left[\text{Error}\right] &< \left(2^B - 1\right) \frac{1}{2} \exp\left(-\frac{B\mathcal{E}_b}{4\sigma^2}\right) & (21)\\ &< 2^B \exp\left(-\frac{B\mathcal{E}_b}{4\sigma^2}\right) & (22)\\ &= \exp\left(B \log_e 2 - \frac{B\mathcal{E}_b}{4\sigma^2}\right) & (23)\\ &= \exp\left(-B\left(\frac{\mathcal{E}_b}{4\sigma^2} - \log_e 2\right)\right). & (24) \end{aligned}$$

So, is the unreliability small for large packet sizes B? The answer depends on how large the energy per bit $\mathcal{E}_b$ we invest in: if it is large enough:

$$\mathcal{E}_b > 4\sigma^2 \log_e 2, \tag{25}$$

then for large values of B the probability of error goes to zero.

Reprise

We have seen a very different type of block communication, position modulation, that is arbitrarily reliable *and* energy efficient. It came about by relaxing the instantaneous energy constraint to an *average* energy constraint. A few key questions arise naturally at this point.

1. In engineering practice, it may not be possible to transmit a large voltage (as the packet size B grows, the voltage magnitude also grows without bound). Indeed, most electronic devices come with both an average *and* peak power rating. If the peak power allowed is finite, the pulse modulation scheme described here will not work anymore. In this case, there are no known simple ways to get energy efficient reliable communication and we will address this issue in the lectures to come.

2. Is there something fundamental about the threshold for energy per bit given in Equation (25)? Perhaps there are other schemes that promise arbitrarily reliable communication and yet allow lower energy per bit than the threshold in Equation (25)?

   (a) We will see in a homework exercise that the threshold in Equation (25) can be lowered by a factor of 2 by doing a more nuanced calculation of the error probability (as compared to the crude union bound used in Equation (14)).

   (b) It turns out that the improved threshold of half of that in Equation (25) is truly fundamental:

   > any communication scheme promising arbitrarily reliable communication over an AWGN channel has to expend energy per bit of *at least* $2\sigma^2 \log_e 2$.

   In this sense $2\sigma^2 \log_e 2$ is a fundamental number for reliable communication over the AWGN channel. We will get more intuition on where this comes from shortly.

Apart from these aspects, position modulation is important on its own right.

- We will see position modulation shows up naturally in *deep space*communication (where data rate is much less an issue than energy efficiency). Deep space communication covers both satellite communication and earth's communication with remote interplanetary space missions.

- It is conjectured (based on experimental data) that the human nervous system communicates using position. Apparently the entire image captured by the human eye needs just three "spikes" (synapses) or so and all the visual information is said to be contained in the *spacing* between the synapses. The book

Spikes: Exploring the Neural Code by Fred Rieke, David Warland, Rob deRuytervanSteveninck, and William Bialek, MIT Press, 1999,

makes for fascinating reading.

Looking Ahead

We have delivered on one of the free food promised earlier: reliable communication in an energy efficient manner. But this still entailed very small data rates. In the next lectures, we will see what it takes to do arbitrarily reliable communication with non-zero rates as well.