6.441 Transmission of Information February 7, 2006

Lecture 1

Lecturer: Madhu Sudan *Scribe: Elena Grigorescu*

1 Administrative Issues

Lecturer: Madhu Sudan, madhu@mit.edu
TA: Chung Chan, chungc@mit.edu
Website: http://theory.csail.mit.edu/∼madhu/ST06

Note: Visit the website in order to sign up for scribing and to fill up the questionnaire (if you didn't do it in class).

Class policy:

- 4 PSets. The first PSet is out today and will be due in about two weeks.
  Collaboration is encouraged, however the write-ups must be done separately and all the sources should be mentioned.
- 1 Midterm.
- 1 Project. The project consists in presenting one of the papers on the website, in teams of two. See more instructions online.
- 1 Scribe notes.

2 General Course Topics:

1. Mathematics of Information transmission
2. How to quantify information
3. How to quantify the 'capacity' of a communication channel
4. How to manipulate these quantities

We will use Probability Theory as a main mathematical tool in defining notions such as **information** and **entropy**.

3 Motivating Scenario

As an introductory example let us start with considering the following scenario. A space satellite is supposed to collect data and send it back to Earth. Let us assume that its sensor measures the surrounding temperature (say for now, as an integer) and the transmitter sends it to Earth at a rate 1 bit/time unit. However, the transmission is not perfectly accurate and therefore the received data is erroneous. The general question that we would like to answer is whether communication is feasible in this kind of settings. Let us model the above problem in a way that would allow us to make mathematical deductions.

3.1 A Simple Model

The Sensor:
Let $x_0, x_1, \ldots, x_t, x_{t+1} \ldots$ denote the measured temperatures at time $0, 1, \ldots$. For simplicity, we assume that each x_i is an integer. Since temperature is a continuous function, we do not expect drastic variations between consecutive measurements. Suppose that the following holds regarding consecutive measurements:

$$Pr[|x_{t+1} - x_t| \geq k] \leq 8^{-k}.$$

It follows that

$x_{t+1} = x_t$ w.p. $\frac{7}{8}$

$x_{t+1} = x_t + 1$ w.p. $\frac{7}{128}$

$x_{t+1} = x_t - 1$ w.p. $\frac{7}{128}$

$\vdots$

Our goal is to be able to transmit x_t for each i. Since we expect $|x_{t+1} - x_t|$ to be small, it is less costly to send $y_t = x_{t+1} - x_t$, rather than x_{t+1} at each time.

Therefore, we will need to send

$y_t = 0$ w.p. $\frac{7}{8}$

$y_t = +1$ w.p $\frac{7}{128}$

$y_t = -1$ w.p. $\frac{7}{128}$

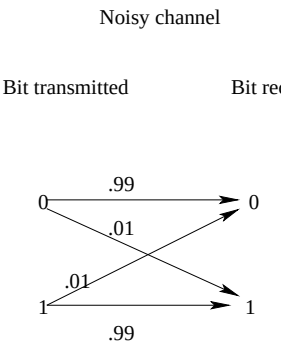
$y_t = +2$ w.p. etc.

$y_t = -2$ w.p. etc.

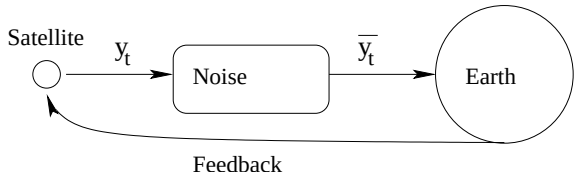
$\vdots$

All this data will be first encoded in some binary form (say, $0 \to 0$, $+1 \to 100$, $-1 \to 101$, $+2 \to 1100$, $-2 \to 1101$, etc.) and then sent through a noisy transmission channel.

The Transmission Channel:
Consider a transmission channel that flips a bit w.p .01, as described in the below figure. We will



moreover assume that the channel's decision to flip a bit at a certain moment in time is independent of its past or future behavior.
Let us first examine in what conditions data transmission across this channel is possible. A first question that we might ask is if we could send each y_i at one unit of time. This obviously cannot be done since the expected channel capacity is less than 1 bit/time, while the expected encoding length is greater than 1.
Instead, we could try and buffer the information that we get for say, 100 units of time and then send it. In analyzing the feasibility of this approach today we will be making crude assumptions and approximations, which however will be improved later in the course.
First, note that under the above premises, in the 100-bit sequence the expected number of 0's is 87 ($=\frac{7}{8}$.) Therefore, in order to specify the location of all these 0's we need $\log \binom{100}{87} \approx 53$ bits.
Similarly, the expected number of ± 1's is ≈ 11, and we need $\log \binom{13}{11} \approx 7$ bits to specify these positions, and 11 more bits to distinguish between $+1$ and -1. In addition, we should consider an expected additional cost of ≤ 3 per y_t.
Summing up, the total cost of about 77 bits seems feasible compared to the channel's capacity of 100 bits. However, we should take into account the fact that the channel is noisy and thus some sort of redundancies should be added in order to correctly decode from transmission error. For now, we will make some more simplifying assumptions about our model. Note that the expected error of the channel is of 1 bit, and assume that the channel actually makes exactly one error. Moreover, assume that the satellite can receive feedback from Earth with full accuracy, such that it can compute the position of the error and send it back to Earth. This will take $\log 100 \approx 7$ more bits. Even so, we are still in the limits of our capacity, which concludes the feasibility of this unrealistic model.



In future lectures we will see that the same results can be obtained however even when we do not make such optimistic assumptions as above.

4 Course highlights

Having introduced the motivating example of our topic, we can now be a bit more specific about the content of the course. The following are topics that we aim to approach.

- Review probability (today)
- Entropy and Information (next few lectures)
- Asymptotic Equipartition Property (the information theorists' Law of Large Numbers)
- Source coding (looks at the rate at which a source is producing information)
- Channel coding (looks at coding channels as the one described today: discrete channels)
- Continuous channels and Gaussian Error
- Network Information Theory (applications in other settings, such as stock markets, gambling, etc.)

References

1. Elements of Information Theory, T. Cover and J. Thomas - available on Reserve at the Barker Library and CSAIL Reading Room.

2. Course website, Scribe notes, Scribbled notes, Lectures notes from previous offerings.

5 Brief Review of Probability Theory

Probability Space: $(\Omega, \mathcal{F}, P)$, where Ω is an underlying ground set, $\mathcal{F}$ is the power set of Ω and P is a probability measure associated with events $E \in \mathcal{F}$.
To each probability space one can associate a **random variable** X distributed according to P. If Ω is a finite set, then $P(x) \geq 0$ for all $x \in \Omega$ and $\sum_{x \in \Omega} P(x) = 1$.
The **expectation** of a real valued random variable X is defined to be $E[X] = \sum_{x \in \Omega} xP(x)$.
The **indicator variable** of an event A is $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise. Thus, $E[1_A] = Pr[A]$.

The following are basic facts that we will be using extensively.

1. $Pr[E_1 \cup E_2] \leq Pr[E_1] + Pr[E_2]$.

2. $E[X_1 + X_2] = E[X_1] + E[X_2]$.

3. For a random variable $X \geq 0$, **Markov's inequality** states

$$Pr[X \geq kE[X]] \leq \frac{1}{k}.$$

Therefore,

$$Pr[(X - E[X])^2 \geq k^2 E[(X - E(X))^2]] \leq \frac{1}{k^2}.$$

The **variance** of X is defined as $Var[X] = E[X^2] - E[X]^2$. As an exercise, show that $Var[X] = E[(X - E(X))^2]$, and thus $Var[X] \geq 0$.

4. The latter inequality can be rewritten in a form known as **Chebychev's inequality**

$$Pr[|X - E[X]| \geq k\sqrt{Var[X]}] \leq \frac{1}{k^2}.$$

By definition, $\sigma[X] = \sqrt{Var[X]}$ is the **standard deviation** of X. As an exercise, figure out when $\sigma[X] = 0$.

5. **Conditional Probabilities:**

$$Pr[E_2|E_1] = \frac{Pr[E_1 \cap E_2]}{Pr[E_1]}.$$

6. One important concept in probability is that of **independence**. Two events E_1 and E_2 are independent if $Pr[E_2|E_1] = Pr[E_2]$.
Consider the following experiment called Random Decreasing Sequence. The sequence is such that, if the random number picked at index i was n_i then at index $i+1$ one picks a random number $n_{i+1} \leq n_i$. The sequence starts at $n_0 = 100$ and ends when $n_t = 1$ for some t. Let E_n be the event that the number n appears in this sequence. We can ask questions such as: What is $Pr[E_{10}]$ or $Pr[E_{11}]$? Are E_{10} and

E_{11} independent? Give these questions some thought...

7. **Chernoff-Hoefding bound:** Given $X_1, \ldots, X_n$ identically and independently distributed (for short, i.i.d.), such that $X_i \in [0, 1]$ and $E[X_i] = \mu$, then

$$Pr[|\frac{\sum_{i=1}^{n} X_i}{n} - \mu| \geq \epsilon] \leq e^{\frac{-\epsilon^2 n}{2}}.$$