

**050.680 — Learning Theory**  
**Problem Set 1**  
**Due 2 February 2007**

**Problem 1**

Let  $\mathcal{L} = \{\mathbb{N} - \{x\} \mid x \in \mathbb{N}\}$ .

- a. Give a learning function that identifies  $\mathcal{L}$ . (Provide a mathematical or algorithmic specification of the learning function if you can, or if you are having trouble explain how it should work in words.)
- b. Consider the following definition:

**Definition.** *A learning function  $f$  is self-monitoring just in case for all texts  $t$ , if  $f$  identifies  $t$ , then (a) there exists a unique  $n \in \mathbb{N}$  such that  $f(\bar{t}_n) = 0$ , and (b) for all  $i > n$ ,  $f(\bar{t}_i) = f(\bar{t}_{i+1})$ .*

Intuitively, a learner is self-monitoring just in case it signals its own successful convergence, where the otherwise useless output 0 signals convergence. Once a self-monitoring learner outputs 0, it can no longer change its mind.

Show that no self-monitoring learning function identifies  $\mathcal{L}$ .

**Problem 2**

Let  $t$  be a text and  $f$  be a learning function.

- a. Show that if  $f$  converges on  $t$ , then  $\{f(\bar{t}_n) \mid n \in \mathbb{N}\}$  is finite.
- b. Why doesn't the converse of (a) hold, i.e., it is not necessarily the case that if  $\{f(\bar{t}_n) \mid n \in \mathbb{N}\}$  is finite, then  $f$  converges on  $t$ .

**Problem 3**

Imagine that you land on a far away island on which there are three tribes which speak three different languages. The language of the first tribe, which we will call  $L_1$ , consists of the single word "Rock", the second one,  $L_2$ , consists of the two words "Rock" and "Scissors", and the third,  $L_3$ , includes the three words "Rock", "Scissors" and "Paper". Since each of these groups is rather hostile to the others, you are well advised to figure out as quickly as possible the language of the group with which you are interacting. You remember that an anthropologist friend of yours, who had once visited this very same island, told you that all members of these tribes had the curious property that they could not help naming any object with which they came into contact, so long as the language they spoke had a word for the object. So, whenever a member of the third tribe encounters a rock, a pair of scissors or a piece of paper, he feels compelled to cry out "Rock", "Scissors" or "Paper", as appropriate. In contrast, members of the second tribe would cry out "Rock" or "Scissors" whenever coming across a rock or a pair of scissors, but would remain silent in the face of a piece of paper. Members of the first tribe remain silent for all objects but rocks.

**Part A:** You come across a native and decide to follow him around for a while in order to figure out the language that he speaks. Give a learning algorithm  $\mathcal{L}$  that will use a collection of observations, each one consisting of a pairing of an object (either rock, scissors or paper) with a verbal response (either "Rock", "Scissors", "Paper" or silence) to figure out which language your companion speaks. You should

specify your algorithm as precisely as possible, in some sort of pseudo-code if possible, and in words if you are having difficulty. And recall that just as in the Gold paradigm, a learning algorithm should map a set of data to a *single* hypothesis.

**Part B:** You suddenly remember that an anthropologist you know had told you that on this island, there were twice as many rocks as there were pairs of scissors, and three times as many pairs of scissors as pieces of paper. Cleverly, you infer that the probability that your newfound native friend will come across a rock in his travels is twice the probability of coming across a pair of scissors, which in turn is three times the probability of coming across a piece of paper. Doing some quick algebra, you conclude that  $\mu(\{\text{rock}\}) = .6$ ,  $\mu(\{\text{scissors}\}) = .3$ ,  $\mu(\{\text{paper}\}) = .1$ . Suppose now that you are faced with the following sequence of observations of your native friend:

SEES	SAYS
rock	“Rock”
paper	silence
rock	“Rock”

You hastily conclude that your friend must speak  $L_1$ . If it turns out that your friend actually speaks  $L_2$ , what is the error set for your hypothesis? If you wanted to achieve 85% accuracy, would this be good enough? That is, is  $\text{error}_\mu(L_2) < .15$ ?

**Part C:** Continue to assume that your friend actually speaks  $L_2$ . Of all the possible sequences of observations of length 3, which ones would incorrectly lead the learning algorithm  $\mathcal{L}$  you gave in part A to an incorrect conclusion? What is the probability that a sequence of 3 observations will be one of these (i.e.,  $\mu^3\{s \in S(3, L_2) | \mathcal{L}(s) = L_2\}$ , where  $S(n, L)$  denotes a sequence of  $n$  observations drawn from  $L$ )? What is the probability of such a misleading sequence in the case of an arbitrary number  $n$  of observations? State how many observations you would need to reach 90% confidence that your hypothesis was correct by finding the least  $m$  for which

$$\mu^m\{s \in S(m, L_2) | \mathcal{L}(s) = L_2\} > .9$$

(The notation  $\mu^m$  refers to the probability distribution of sequences of observations of length  $m$ . Since the observations are IID, this can be computed as the product of the individual observation probabilities under distribution  $\mu$ .) Suppose that you only require that your hypothesis be 85% accurate. Determine the number of observations that would now be required to obtain 90% confidence by finding the least  $m$  for which

$$\mu^m\{s \in S(m, L_2) | \text{error}_\mu(\mathcal{L}(s)) < .15\} > .9$$

**Part D:** Suppose that your friend speaks  $L_3$ . What is the probability that a sequence of  $n$  observations will lead your learner to an incorrect conclusion (i.e.,  $\mu^n\{s \in S(n, L_3) | \mathcal{L}(s) = L_3\}$ )? In this case, how many observations would be required to reach 90% confidence that your hypothesis is correct, i.e., what is the least  $m$  for which

$$\mu^m\{s \in S(m, L_3) | \mathcal{L}(s) = L_3\} > .9$$

What happens in this case to the number of observations  $m$  required for 90% confidence if you were to only require 85% accuracy? That is, what is the least  $m$  for which

$$\mu^m\{s \in S(m, L_3) | \text{error}_\mu(\mathcal{L}(s)) < .15\} > .9$$