Problem set 1
Some probability via Hilbert space.

Math 212a14

Sept. 4, 2012, Due Sept. 16

This is a rather long problem set dealing with a chunk of probability theory that we can do in Hilbert space terms (without fully developing measure theory). But it shouldn't take you more than three hours to do. Please do not leave it to the last minute and then run out of time. There are eleven problems in all.

Contents

1 The L_2 Martingale Convergence Theorem.

Let $\mathbf{H}$ be a Hilbert space and let $\{x_n\}$ be a sequence of elements of $\mathbf{H}$ which satisfy the following two conditions:

- The $\|x_n\|$ are bounded, i.e. there is a constant M such that

$$\|x_n\| \leq M \quad \text{for all } n$$

and

- For all pairs m and n with $n > m$

$$(x_n - x_m, x_m) = 0. \tag{1}$$

1. Show that the x_n converge to an element x and that $\|x\| \leq M$. In fact, show that $\|x\| = L$ where L is the greatest lower bound of the M in the first condition.

Once you have figured out this problem, you will agree that the proof is quite transparent. You may wonder about the fancy title. In fact, this result is (part of, or a version of) a key result in probability theory which unifies many so called "limit theorems". It is also a very good entry into getting an intuitive understanding of these theorems. The appropriate language for stating these theorems is measure theory, and therefore any advanced book on probability theory either assumes that the reader has had a course on measure theory, or spends the early chapters on an exposition of measure theory. The martingale convergence theorem usually does not make its appearance until the last third of the book. I plan to get into measure theory in the next few weeks, but in this problem set I want to develop a good bit of the language and some of the results of probability theory using Hilbert space tools.

Here is an example of an "abstract" limit theorem which is in fact a consequence of the L_2 martingale convergence theorem, but is easier to prove directly:

2. Let $y_k \in \mathbf{H}$ be such that

$$(y_i, y_j) = 0 \quad i \neq j$$

and such that there is a constant K such that

$$\|y_j\|^2 \leq K$$

for all j. Show that for any real number $s > \frac{1}{2}$ we have

$$\lim_{n\to\infty} \frac{1}{n^s}(y_1 + \cdots + y_n) \to 0.$$

A matter of great importance is to understand what happens (under additional hypotheses) at the critical value $s = \frac{1}{2}$. We will deal with this in a later problem set where we will study the "central limit theorem".

To understand why Problem **2** is in fact a consequence of Problem **1**, consider the case $s = 1$ and take

$$x_n := \sum_1^n \frac{1}{k} y_k.$$

Then the $\|x_n\|$ are bounded because $\sum 1/k^2$ converges, and condition (1) clearly holds. We have

$$y_k = k(x_k - x_{k-1})$$

so

$$y_1 + \cdots + y_n = \sum_1^n k(x_k - x_{k-1}) = nx_n - \sum_1^n x_{k-1}.$$

Dividing by n gives

$$\frac{1}{n}(y_1 + \cdots + y_n) = x_n - \frac{1}{n}\sum_1^n x_{k-1}.$$

Since we know that $x_n \to x$ we also know that the "Cesaro average" $(1/n)\sum_1^n x_{k-1}$ approaches x so the right hand side tends to $x - x = 0$. The same "summation by parts" argument works for any $s > \frac{1}{2}$ to derive Problem **2** from Problem **1** but is a little harder to write down. The direct proof is easier.

2 Probabilistic language in Hilbert space terms.

In what follows we will take $\mathbf{H}$ to be $L_2([0,1])$, the completion of the space of continuous functions on the unit interval relative to the metric determined by the scalar product

$$(f, g) = \int_0^1 f(t)\overline{g(t)}dt.$$

(There is, of course, no difference between $L_2(\mathbf{T})$ and $L_2([0,1])$, but since the convention is to let probabilities range between 0 and 1, we will take $\mathbf{H} = L_2([0,1])$.)

The space $L_2([0,1])$ has certain additional structures, beyond merely being a Hilbert space. Once we have extracted the necessary additional structures we could assemble them into a collection of axioms for what we might call a "probability Hilbert space".

2.1 Expectation.

The space $\mathbf{H}$ has a preferred element $\mathbf{1}$ which we might think of as the function which is identically one. More precisely, it is the linear function on $\mathcal{C}([0,1])$ which assigns to any continuous function its integral over the unit interval:

$$f \mapsto \int_0^1 f(t)dt = (f, \mathbf{1}).$$

We have

$$(\mathbf{1}, \mathbf{1}) = 1.$$

We make the definition

$$E(X) := (X, \mathbf{1}) \quad \text{for any } X \in \mathbf{H}$$

and call $E(X)$ the **expectation** of X. We will call an arbitrary element of $\mathbf{H}$ a **random variable**. (Strictly speaking we should call these "square integrable random variables", but we won't be considering any other kind in this problem set.)

Suppose that $A \subset [0,1]$ is an interval, or a finite union of intervals. The linear function which assigns to any continuous function its integral over A,

$$f \mapsto \int_A f(t)dt$$

is continuous in the $\|\cdot\|_2$ norm:

$$\left|\int_A f(t)dt\right| \leq \int_A |f(t)|dt \leq \int_{[0,1]} |f(t)|dt = (|f|, \mathbf{1}) \leq \|f\|_2 \|\mathbf{1}\|_2.$$

so it is given by scalar product with an some element of $\mathbf{H}$ which we shall denote by $\mathbf{1}_A$. The meaning of this notation is that we can "represent" this element by the indicator function of the set A, i.e. by the function which is one on A and zero on the complement of A. Indeed, we can find a sequence of continuous functions which converge to this indicator function in the L_2 norm. There is little danger in thinking of $\mathbf{1}_A$ as being an actual function, or as representing the set A, provided that we understand that modifying the set A by inserting or removing a finite number of points (or, as we shall see later, by a set of measure zero) gives the same element $\mathbf{1}_A \in \mathbf{H}$. Clearly

$$E(\mathbf{1}_A) = (\mathbf{1}_A, \mathbf{1}) = (\mathbf{1}_A, \mathbf{1}_A) = \|\mathbf{1}_A\|^2$$

is just the sum of the lengths of the disjoint intervals which comprise A. (It requires a little combinatorial lemma at this point to prove that this value does not depend on how we break A up into a union of disjoint intervals.) We will denote this value by $P(A)$ and call it the **probability** of the **event** A. So

$$P(A) := E(\mathbf{1}_A) = \|\mathbf{1}_A\|^2. \tag{2}$$

2.2 Conditional probability is a Fourier coefficient.

Suppose that $\mathbf{1}_A \neq 0$, and let π_A denote orthogonal projection onto the one dimensional space spanned by $\mathbf{1}_A$. Let B be some other finite union of intervals, so that $A \cap B$ is again a finite union of intervals.

3. Show that

$$\pi_A(\mathbf{1}_B) = \frac{P(A \cap B)}{P(A)} \mathbf{1}_A.$$

In elementary probability theory the expression $\frac{P(A \cap B)}{P(A)}$ is known as the "conditional probability of B given A" and is denoted by $P(B|A)$,

$$P(B|A) := \frac{P(A \cap B)}{P(A)}.$$

So we can write

$$\pi_A(\mathbf{1}_B) = P(B|A)\mathbf{1}_A. \tag{3}$$

Later on in this problem set, we will enlarge our set of "events" and take (2) and (3) as *definitions*. For this we need to make use of some additional structure of $L_2([0,1])$.

2.3 Reality.

If $\{g_n\}$ is a Cauchy sequence (in the $\|\cdot\|_2$ norm) of elements of $\mathcal{C}([0,1])$ then so is the sequence $\{\overline{g_n}\}$, where, of course,

$$(\overline{f})(t) := \overline{f(t)}.$$

This implies that the complex conjugation operation extends to $\mathbf{H}$ as an anti-linear map of $\mathbf{H}$ onto itself, and $\|\overline{g}\|_2 = \|g\|_2$. Thus we can define $\mathbf{H_R}$ to consist of those $g \in \mathbf{H}$ which satisfy $\overline{g} = g$. It is a vector space over the real numbers, and it is easy to check that every such element is the limit of a Cauchy sequence of real continuous functions. The scalar product of any two elements of $\mathbf{H_R}$ is real.

Alternatively, we could have developed the whole theory of Hilbert spaces over the real numbers, and then obtained $\mathbf{H_R}$ as the completion of the real valued continuous functions on $[0,1]$ under the $\|\cdot\|_2$ norm.

2.4 Covariance and Variance.

Let X and Y be real valued random variables. Define the **covariance** of X and Y by

$$\mathrm{cov}(X,Y) := (X - E(X)\mathbf{1}, Y - E(Y)\mathbf{1}) = (X,Y) - E(X)E(Y).$$

Define the **variance** of X by

$$\mathrm{var}(X) := \mathrm{cov}(X, X) = (X, X) - E(X)^2.$$

We say that X and Y are **uncorrelated** if $\mathrm{cov}(X, Y) = 0$. Notice that if $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$ are events, then X and Y are uncorrelated if and only if $P(A \cap B) = P(A)P(B)$ or (if $P(A) \neq 0$) if $P(B|A) = P(B)$. We say that the events are **independent**. For more general random variables, we will introduce a notion of independence which is much more restrictive than being uncorrelated.

2.5 The L_2 law of large numbers.

Suppose $\{Z_i\}$ is a family of pairwise uncorrelated random variables whose variances are uniformly bounded:

$$\mathrm{var}(Z_i) \leq K$$

for some constant K for all i. We can then apply Problem **2** to $y_i := Z_i - E(Z_i)\mathbf{1}$ to conclude that

$$\frac{1}{n}(Z_1 + \cdots + Z_n) - \frac{1}{n}\left(E(Z_1) + \cdots + E(Z_n)\right)\mathbf{1} \to 0.$$

For example, if all the $E(Z_i)$ are equal, say equal to some number m, then

$$\frac{1}{n}(Z_1 + \cdots + Z_n) \to m\mathbf{1}.$$

In this equation, convergence is in the sense of L_2. It will require more work for us to formulate and prove a corresponding result where convergence is "almost everywhere". We will do this after developing the machinery of measure theory - we will then also be able to formulate and prove an "almost everywhere" version of the martingale convergence theorem.

But all versions of the law of large numbers are meant to justify the identification of the intuitive notion of probability with that of long term behavior.

We are going to greatly enlarge the class of "events" and this is going to take a lot of technical work. The idea is this: let S be any subset of $[0, 1]$. We can consider the function 1_S where $1_S(x) = 1$ for $x \in S$ and $1_S(x) = 0$ for $x \notin S$. Conversely, let f be any function on $[0, 1]$ which takes on only the values 1 and 0, then we can let S be the set of x where $f(x) = 1$, and then $f = 1_S$. Now a function f which takes on only the values 0 and 1 can be characterized as follows:

$$0 \leq f(x) \leq 1 \quad \forall x,$$

$$\max\{f(x), 1 - f(x)\} = 1, \quad \forall x,$$

and

$$\min\{f(x), 1 - f(x)\} = 0 \quad \forall x.$$

Since the elements of $\mathbf{H_R}$ are not functions, we can not use this characterization of the indicator function of a set directly. The trouble is with the quantifier $\forall x$.

Changing "values" at a set of measure zero does not have any effect on an element of $\mathbf{H_R}$. So what we need is to show that there is a notion of partial order $\preceq$ on $\mathbf{H_R}$ which is compatable with the Hilbert space structure and which is a substitute for the order relation $f(x) \leq g(x)\ \forall x$. Before embarking on this technical stuff, I thought it might be useful to develop some probablistic intuition for the concepts we are introducing.

3 Simulation.

The m.file randomwalk.m in the mfiles (MATLAB) folder simulates a one dimensional random walk with N steps. For convenience, I reproduce it here:

1. % plots a random walk of length N
2. close all
3. clc
4. N=input('N= ')
5. Q=zeros(1,N);
6. R=rand(1,N);
7. Q=Q+(R>.5);
8. P=2*Q-1;
9. W=[0 cumsum(P)];
10. plot(W)

Step 1 is not a command at all, and is just there to describe what the m.file does. Steps 2 and 3 clear the display and command window and Step 4 calls for the input of the number of random walk steps desired. Step 5 is just an initialization. The key steps are 6 and 7. The MATLAB command rand(M,N) produces a M $\times$ N matrix each of whose entries is a number "chosen at random" between 0 and 1, and all the matrix entries chosen completely independently of one another.

There is obviously a deep paradox involved in how a purely mechanical device like a computer can produce numbers "chosen at random". I will leave this issue for courses in the CS department or the philosophy department. Perhaps also the theology department. I will just take this computer capacity for granted.

In step 7, the entry R>.5 produces a matrix the same size as the matrix R, whose entries are 1 or 0 according as the corresponding entry of R is or is not >.5. In other words, it has the effect of applying the function $\mathbf{1}_{[.5,1]}$ to each of the entries of R. The net effect of step 7 is to produce a vector of length N whose entries are 0 or 1, each occurring with probability one-half. Step 8 converts this into a vector whose entries are ± 1 each with probability one-half.

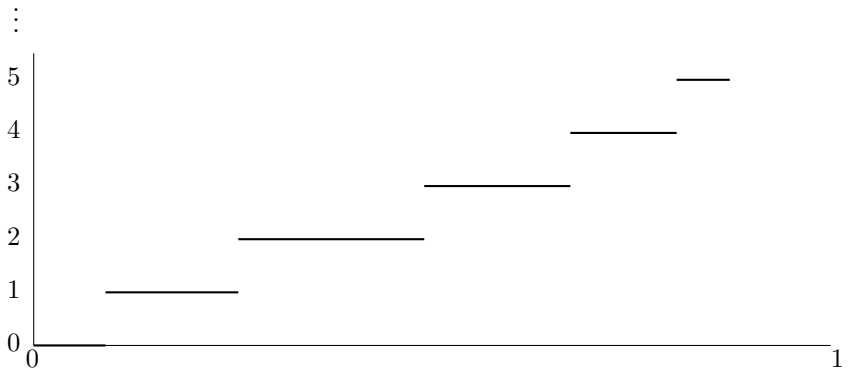

Figure 1: The graph of function f such that $f(\text{rand})$ simulates an experiment which has outcome i with probabiity p_i. The length of the $i-$th interval of constancy is p_i.

Step 9 produces a vector whose first entry is 0 and whose next entries are the cumulative sums of the entries of P and step 10 plots the entries of W against the index of the entry, i.e plots the points (i,W(i)) and joins these points by straight line segments.

If you have access to Matlab you should run this program for various values of N, including some large N to get a feel for what is going on. Otherwise run a similar program in whatever platform you prefer. The intuition obtained will be useful when we study Brownian motion and Weiner measure later in the semester.

Suppose we wanted to simulate an unfair coin where the probability of heads is p, not necessarily .5. We could then apply the function $\mathbf{1}_{[1-p,1]}$ to each of entries of R. This will produce a matrix whose entries are 0 or 1, with 0 occurring with probability $1-p$ and 1 with probability p.

More generally, suppose we want to simulate an experiment whose outcome can be any integer from 0 to n, where the probability of getting i is p_i. Then all we have to do is construct a "step function" whose values are these integers, where the value i is taken on on an interval of length p_i. For example, the function f graphed in Figure 1 will do. Then applying f to each entry of R will produce a simluation of independent trials of the experiment, one trial for each position of R.

In general, if we want to simulate a random variable which takes on non-negative integer values, we don't have to restrict ourselves to the function graphed in Figure 1. In fact, let g be any function defined on $[0,1]$ which takes on non-negative integer values, and suppose (for the sake of simplicity) that for each non-negative integer k, the set

$$g^{-1}(k)$$

is a finite union of intervals. (An interval can be open (without end points), closed (with both end points) or half open (with one end point as in Figure 1).) Suppose that

$$\mu[g^{-1}(k)] = p_k$$

Here $\mu(S)$ denotes the "measure " of a set S. If S is a disjoint union of intervals, then $\mu(S)$ is the sum of the lengths of constituent intervals. If g satisfies this condition, then clearly

$$X = g(\text{rand})$$

simulates an experiment with

$$\text{Prob}(X = i) = p_i, \quad i = 0, 1, 2, \ldots, \quad \sum p_i = 1.$$

So the function f in Figure 1 is just one of many. It is characterized by being monotone (non-decreasing). Any function g that we choose belongs to (more precisely "represents") an element of $\mathbf{H_R}$, and this is why we call elements of $\mathbf{H_R}$ random variables.

We can allow for the possibility that any non-negative integer value be achieved. The condition that a g simulating such an experiment belong to $\mathbf{H_R}$ is that

$$\sum i^2 p_i < \infty.$$

Thus our space $\mathbf{H_R}$ will not simulate *all* random experiments, only the square integrable ones.

3.1 The Poisson distribution.

This describes an experiment whose outcome can be any non-negative integer, and where the probablity of having the outcome i is

$$p_i(\lambda) = \frac{\lambda^i}{i!} e^{-\lambda}.$$

Here λ is any non-negative real number. (In the mathematics literature, λ is a standard name for this parameter, but MATLAB prefers the English alphabet, so we will also use letters such as r or p for this parameter.)

4. Let X be a random variable which takes on only non-negative integer values i with probablity $p_i(\lambda)$ as given above. Show that

$$E(X) = \lambda \quad \text{and} \quad \text{var}(X) = \lambda.$$

[A neat way of doing this is (for any random variable X that takes on only non-negative integer values n, with probability p_n) is to introduce the **generating function**

$$p_X(z) := \sum_0^\infty p_n z^n$$

and then to evaluate $p'(1)$ and $p''(1)$. Express the expectation and the variance in terms of these derivatives and then apply the result to the Poisson distribution.]

As mentioned above, many different random variables can simulate the same experiment, but it will be convenient for us to choose the monotone step function as a standard simulator. So we let Y_λ denote the monotone step function with values in the non-negative integers, and where the interval where $Y_\lambda = i$ has length $p_i(\lambda)$.

The m.file poisson.m calls for the input of the parameter r ($= \lambda$) and an integer N, and produces a row vector Q of size N whose entries are non-negative integers chosen independently according to the Poisson law. Here is the m.file:

```
1. %produces a vector Q of size N with integer entries
2. %distributed according to the Poisson distribution
3. N=input('N= ')
4. r=input('r= ')
5. p=exp(-r);
6. F=p;
7. i=1;
8. Q=zeros(1,N);
9. R=rand(1,N);
10. while p > eps;
11. Q=Q+(R > F);
12. p=(r/i)*p;
13. F=F+p;
14. i=i+1;
15. end
```

It works by first checking whether any entry of R is $> p_0(r) = e^{-r}$. If so, it adds 1 to the corresponding entry of Q. Then it checks whether the entry is $> p_0(r) + p_1(r)$. If so, it adds another 1 to the corresponding entry of Q etc. It computes $p_i(r)$ as it goes along. We do not want the machine to continue once i is so large that $p_i(r)$ is below the tolerance that MATLAB uses in its computations. This is the meaning of line 10: the command "eps" returns the value of the machine accuracy, and we want the procedure to stop when $p_i(r)$ gets below this value.

There will be problems with this program if r is so large that $e^{-r} < eps$ for then the program can't even get started. Also, this program is not the most efficient when the parameter r is only moderately large, since it spends a lot of time checking events of low probability, but it is fine for low values of r and useful for our theoretical discussion.

Notice that this program is a Matlab implementation of repeated independent trials of the elment $f \in \mathbf{H_R}$ whose graph is given by Figure 1 with $p_i = e^{-\lambda}\frac{\lambda^i}{i!}$. You might want to check the law of large numbers by counting how many zeros, ones, twos etc. occur in Q for large N.

3.2 Comparing Poisson and Bernoulli.

We will let B_p denote the monotone step function which is equal to 0 on the interval $[0, 1-p)$ and equal to 1 on the interval $[1-p, 1]$. The letter B stands for Bernoulli, who considered random variables which could take on only the values 0 and 1, and takes on the value 1 with probability p. Of course B_p is only one of many such random variables, but all have the same expectation and variance, namely

$$E(B_p) = p \quad \text{and} \quad \text{var}(B_p) = p - p^2.$$

We also note the following useful fact:

$$e^{-p} = 1 - p + \frac{p^2}{2!} - \frac{p^3}{3!} + \cdots$$

is (for $0 \leq p \leq 1$) an alternating series with terms decreasing in absolute value, and so $e^{-p} - 1 + p > 0$ or

$$1 - p < e^{-p}.$$

This says that the interval where $B_p = 0$ is shorter than the interval where $Y_p = 0$.

5. Show that the set $A := \{x \in [0,1] | B_p(x) \neq Y_p(x)\}$ consists of two intervals, and

$$P(A) \leq p^2.$$

We can write this result in the more convenient shorter notation

$$P(B_p \neq Y_p) \leq p^2.$$

3.3 Distance in law.

We continue to discuss the situation where our random variables take on only integer values, and the inverse image of each integer is a finite union of intervals. Neither of these hypotheses is essential, but I want to develop some intuition and these assumptions provide a useful crutch.

Each such random variable X assigns a non-negative weight or probability p_k to each integer k where the p_k are non-negative and $\sum_k p_k = 1$. Suppose that Y is a second non-negative integer valued random variable which has the probability assignments $P(Y = k) = q_k$. We say that X and Y "have the same law" or are "equal in law" if $p_k = q_k$ for all k. More generally we might want

to measure how far the overall distribution of the p's is from that of the q's by defining

$$d_{\text{law}}(X,Y) \stackrel{\text{def}}{=} \sup_A |\text{Prob}(X \in A) - \text{Prob}(Y \in A)| \tag{4}$$

where the supremum is taken over all subsets A of the integers.

Notice that if $d_{\text{law}}(X,Y) = 0$ then X and Y are equal in law, and if X, Y, Z are three random variables then for any (measurable) set A we have

$$\begin{aligned} |\text{Prob}(X \in A) - \text{Prob}(Z \in A)| &= |\text{Prob}(X \in A) - \text{Prob}(Y \in A) \\ &\qquad +\text{Prob}(Y \in A) - \text{Prob}(Z \in A)| \\ &\leq |\text{Prob}(X \in A) - \text{Prob}(Y \in A)| \\ &\qquad +|\text{Prob}(Y \in A) - \text{Prob}(Z \in A)| \\ &\leq d_{\text{law}}(X,Y) + d_{\text{law}}(Y,Z). \end{aligned}$$

Taking the supremum over all A gives the triangle inequality

$$d_{\text{law}}(X,Z) \leq d_{\text{law}}(X,Y) + d_{\text{law}}(Y,Z).$$

So "distance in law" satisfies the conditions for a pseudometric. But different random variables can be equal in law, so saying that two random variables are equal in law definitely does *not* imply that they are equal, as we have seen.

More generally, suppose that X and Y are discrete random variables with possible values $\{x_k\}$ and $\{y_k\}$

$$\text{Prob}(X = x_k) = r_k, \quad \text{Prob}(Y = y_k) = s_k$$

so that

$$\sum_k r_k = \sum_k s_k = 1$$

when the sums are taken over all k. By throwing in 0's for r's or s's corresponding to values which don't occur in X or Y and relabeling, we can assume that the sets $\{x_k\}$ and $\{y_k\}$ are the same. Now when we measure distance in law we let A range over all subsets of these possible values.

Let B_+ denote the set of all x_k for which $r_k \geq s_k$ and B_- the complement, i.e. the set where $r_k < s_k$. In words, B_- is the set of values where Y has a larger probability of occuring than X. For any set A we have the disjoint union

$$A = (A \cap B_+) \cup (A \cap B_-)$$

and so

$$\begin{aligned} \text{Prob}(X \in A) - \text{Prob}(Y \in A) &= \text{Prob}(X \in A \cap B_+) - \text{Prob}(Y \in A \cap B_+) \\ &\qquad +\text{Prob}(X \in A \cap B_-) - \text{Prob}(Y \in A \cap B_-) \\ &= \sum_{x_k \in (A \cap B_+)} (r_k - s_k) + \sum_{x_k \in (A \cap B_-)} (r_k - s_k). \end{aligned}$$

The summands in the first sum are all non-negative, and in the second sum are all negative. So there will be cancellation between these two sums unless one or the other vanishes. So in taking the supremum in the definition (4) we may assume that $A \subset B_+$ or $A \subset B_-$, and that, in fact $A = B_+$ or $A = B_-$. But since

$$\text{Prob}(X \in B_+) + \text{Prob}(X \in B_-) = 1$$

and similarly for Y, we have

$$\text{Prob}(X \in B_+) - \text{Prob}(Y \in B_+) = -[\text{Prob}(X \in B_-) - \text{Prob}(Y \in B_-)$$

so the supremum in (4) is achieved by taking $A = B_+$ or $A = B_-$. Since these two sums are equal, their common value is equal to one half of their sum and so we get

$$d_{\text{law}}(X, Y) = \frac{1}{2} \sum_k |r_k - s_k| \tag{5}$$

where the sum is over all k. so this is an alternative definition for the distance in law between two discrete random variables.

Let f be any function satisfying

$$|f(x)| \leq 1$$

for all x. Then

$$|E(f(X)) - E(f(Y))| = |\sum_{x_k \in B_+} f(x_k)(r_k - s_k) + \sum_{x_k \in B_-} f(x_k)(r_k - s_k)|.$$

The right hand side of this expression will be maximized (subject to the constraint $|f(x)| \leq 1$) if we choose $f \equiv 1$ on B_+ and $f \equiv -1$ on B_- in which case the expression on the right becomes

$$\sum_k |r_k - s_k| = 2d_{\text{law}}(X, Y).$$

So if we define

$$d_f(X, Y) = \frac{1}{2}|E(f(X)) - E(f(Y))|, \tag{6}$$

we have

$$d_{\text{law}}(X, Y) = \sup_{|f| \leq 1} d_f(X, Y). \tag{7}$$

Notice that the definition (6) is given in purely Hilbert space terms.

Let us assume that the set $X \neq Y$ is an event; for example that this subset is a finite union of intervals as in Exercise **5**. (Later on in these notes we shall generalize the notion of an "event" and will see that we have enough events so that $X \neq Y$ is always an event.)

Lemma 3.1 *Let X and Y be any two random variables. Then*

$$d_{\text{law}}(X, Y) \leq \text{Prob}(X \neq Y). \tag{8}$$

Proof. By definition, for any set A,

$$\text{Prob}(X \in A) = \mu[X^{-1}(A)]$$

where $X^{-1}(A)$ is the subset of the unit interval where the function X takes values in A, and $\mu[X^{-1}(A)]$ is the measure (total length) of this subset. So

$$\text{Prob}(X \in A) - \text{Prob}(Y \in A) = \mu[X^{-1}(A)] - \mu[Y^{-1}(A)].$$

Let $\{X = Y\}$ denote the subset of the unit interval where the functions X and Y take on the same value:

$$\{X = Y\} = \{x | X(x) = Y(x)\}.$$

Its complement is the set $\{X \neq Y\}$ where the functions X and Y take on unequal values, and the right hand side of (8) is $\mu[\{X \neq Y\}]$. For any set A we have the disjoint union

$$X^{-1}(A) = (X^{-1}(A) \cap \{X = Y\}) \cup (X^{-1}(A) \cap \{X \neq Y\})$$

and hence

$$\text{Prob}(X \in A) = \mu[X^{-1}(A) \cap \{X = Y\}] + \mu[X^{-1}(A) \cap \{X \neq Y\}]$$

with a similar expression for Y. But clearly we have the equality of the two sets

$$X^{-1}(A) \cap \{X = Y\} = Y^{-1}(A) \cap \{X = Y\}$$

and so their measures are the same. So

$$\text{Prob}(X \in A) - \text{Prob}(Y \in A) = \mu[X^{-1}(A) \cap \{X \neq Y\}] - \mu[Y^{-1}(A) \cap \{X \neq Y\}].$$

The expression on the right lies between $-\mu[\{X \neq Y\}]$ and $+\mu[\{X \neq Y\}]$ so

$$|\text{Prob}(X \in A) - \text{Prob}(Y \in A)| \leq \text{Prob}(X \neq Y).$$

Taking the supremum over all A proves (8).

As an example of the use of this lemma we conclude from Problem **5** that

$$d_{\text{law}}(B_p, Y_p) \leq p^2.$$

Notice that the proof of this inequality depended on a specific choice of random variables B_p and Y_p. But distance in law does not depend on this choice: the triangle inequality allow us to replace B_p by any Bernoulli random variable with parameter p and Y_p by any Poisson random variable with parameter p and to conclude

Lemma 3.2 *Let X_p be a Bernoulli random variable with parameter p and let W_p be a Poisson random variable with the same parameter, p. Then*

$$d_{\text{law}}(X_p, W_p) \leq p^2. \tag{9}$$

For discrete random variables we have a simple definition of "independence". We say that X and Y are independent if for any pair of possible values x of X and y of Y we have

$$P((X = x) \cap (Y = y)) = P(X = x)P(Y = y),$$

in other words, the events $X = x$ and $Y = y$ are independent. It is instructive to write this as

$$P((X, Y) = (x, y)) = P((X = x))P((Y = y)).$$

If X and Y are (discrete) random variables, then $X + Y$ is a again a (discrete) random variable and we can consider the expectation of $f(X + Y)$ where f is a function. If X and Y are independent,

$$E(f(X + Y)) = \sum_{x,y} f(x + y)p_x q_y$$

in the obvious notation. We can write this double sum as an iterated sum

$$E(f(X + Y)) = \sum_x p_x \sum_y f(x + y)q_y = \sum_x p_x E_Y(f(x + Y)).$$

Here, for fixed x we are thinking of $f(x + y)$ as a function of y and taking the expectation with respect to Y. Here is even better notation, we form the function $E_Y(f(\cdot + Y)$ which sends $x \mapsto E_Y(f(x + Y))$ and then take the expectation of this with respect to X:

$$E(f(X + Y)) = E_X[E_Y(f(X + Y))].$$

Here is a lemma which says that adding an independent random variable "fuzzes out" differences in law:

6. Let X, Y, Z be three random variables with Z independent of X and of Y. Then

$$d_{\text{law}}(X + Z, Y + Z) \leq d_{\text{law}}(X, Y). \tag{10}$$

[Hint: Use the definition of distance in law which involves minimizing over functions f.]

7. Show that if Z and W are independent Poisson random variables with parameters λ and μ respectively, then $Z + W$ is a Poisson random variable with parameter $\lambda + \mu$. [Hint: A neat way of doing this is to show that if X and Y are independent non-negative integer valued random variables then we have the relation $p_{X+Y}(z) = p_X(z)p_Y(z)$ for their generating functions.]

3.4 Poisson's law of small numbers.

Here is one version:

Theorem 3.1 *Let $X_1, \dots, X_n$ be independent Bernoullis with parameters p_i. Let*

$$\lambda = p_1 + \cdots + p_n,$$

and let Z_λ be Poisson with parameter λ. Then

$$d_{\text{law}}(X_1 + \cdots + X_n, Z_\lambda) \leq p_1^2 + \cdots + p_n^2. \tag{11}$$

For example, suppose that all the p_i are equal, and so equal λ/n. Then the right hand side of (11) is also equal to λ/n. So for λ fixed and n large, this says that a sum of many identical Bernoullis whose expectations add up to finite number λ, can be appoximated (in law) by a Poisson with parameter λ.

The theorem is a bit more general, in that we don't need to have the p_i equal to estimate the right hand side of (11), it is enough that they all be roughly of size λ/n.

Proof of the theorem. Choose Poisson random variables Y_i with parameters p_i which are independent of all the X_i and of each other. Then

$$\begin{aligned} d_{\text{law}}(X_1 + \cdots + X_n, X_1 + \cdots + X_{n-1} + Y_n) &\leq d_{\text{law}}(X_n, Y_n) \\ &\leq p_n^2. \end{aligned}$$

The first inequality is just (10) with $Z = X_1 + \cdots X_{n-1}$ and the second inequality is (9). Next we replace X_{n-1} by Y_{n-1} via

$$\begin{aligned} d_{\text{law}}(X_1 + \cdots + X_{n-1} + Y_n, X_1 + \cdots + X_{n-2} + Y_{n-1} + Y_n) &\leq d_{\text{law}}(X_{n-1}, Y_{n-1}) \\ &\leq p_{n-1}^2 \end{aligned}$$

where the first inequality is (10) with $Z = X_1 + \cdots X_{n-2} + Y_n$ and the second inequality is (9). Proceeding in this way we end up, via the triangle inequality with

$$d_{\text{law}}(X_1 + \cdots + X_n, Y_1 + \cdots + Y_n) \leq p_1^2 + \cdots + p_n^2.$$

But the sum of the Y's is Poisson with parameter $\lambda = p_1 + \cdots + p_n$ by the proposition. QED

There are several important improvements on this theorem, whose proofs are a bit harder, and so we will content ourselves with the statements, postponing or referring the proof to a course on probability theory. The first has to do with the estimate in the theorem. If λ is a large number, we can do better. We can replace (11) by

$$d_{\text{law}}(X_1 + \cdots + X_n, Z_\lambda) \leq 2\frac{1 - e^{-\lambda}}{\lambda} \sum_i p_i^2. \tag{12}$$

Another improvement has to do with relaxing the condition of independence. To illustrate, consider the *birthday problem*: Suppose that we have n people and

we want to compute the probability that at least one pair of the group have the same birthday. Let us make the approximation that there are $c = 365$ days in the year and that birthdays are independentantly distributed with probability $p = 1/c$. Then the probability of no two peaple having the same birthday is clearly

$$\prod_{k=1}^{n-1}\left(\frac{c-k}{c}\right) = \prod_{k=1}^{n-1}\left(1-\frac{k}{c}\right).$$

If we make the approximations

$$\ln(1-\frac{k}{c}) \doteq -\frac{k}{c}, \quad 1 \leq k \leq n$$

the product becomes approximated by

$$\exp(-\frac{(n-1)(n-2)}{2c}).$$

Assuming that n is small with respect to c this can be approximated by

$$e^{-\frac{n^2}{2c}}. \tag{13}$$

This suggests that the probability of getting k matching pairs of birthdays should be approximately Poisson with parameter

$$\lambda \doteq \frac{n^2}{2c} \doteq \frac{n(n-1)}{2c}.$$

This is indeed the case for n moderately large but relatively small in comparison to c. To relate this to our theorem, let $\mathcal{E}$ denote the set of pairs of individuals, so an element $e \in \mathcal{E}$ is a pair, $e = \{x, y\}$ of individuals. For each such $e \in \mathcal{E}$, let X_e denote the random variable which is one or zero according to whether x and y do or do not have the same birthday. Clearly X_e is a Bernoulli random variable, and our assumptions are that

$$\text{Prob}(X_e = 1) = \frac{1}{c}.$$

The sum

$$\sum_{e\in\mathcal{E}} X_e$$

gives the total number of matching pairs, and there are $\frac{n(n-1)}{2}$ pairs, so we would *like* to be able to say that this sum is approximately Poisson distributed with parameter λ as given above. But the theorem does not quite apply, since the random variables X_e are not independent: If x and y have the same birthday, and y and z have the same birthday, then so do x and z. On the other hand, we might expect that this dependency is a second order effect which does not affect the validity of the conlusion of the theorem. We will not go into these more refined versions; again we refer them to a course in probability.

We now interupt this discussion of simulation and return to $\mathbf{H_R}$.

4 Order.

It makes sense to say that a real valued function is non-negative:

$$\phi \succeq 0 \quad \Leftrightarrow \quad \phi(t) \geq 0 \text{ for all } t.$$

We have two natural candidates for how to extend this notion to $\mathbf{H_R}$ and we want to check that these two extended definitions are in fact the same. We could define

1. $g \succeq 0$ if $(g, \phi) \geq 0$ for all $\phi \succeq 0 \in \mathcal{C}([0,1])$ or we could define
2. $g \succeq 0$ if there is a sequence of $g_n \in \mathcal{C}([0,1])$ with $g_n \succeq 0$ and $g_n \to g$.

Clearly $2 \Rightarrow 1$, since $(g, \phi) = \lim(g_n, \phi) \geq 0$. We must show the reverse implication.

For any real number a define

$$a^+ = \frac{1}{2}(a + |a|) = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{if } a \leq 0 \end{cases}.$$

By examining the four possibilities for the signs of a and b we check that

$$|a^+ - b^+| \leq |a - b|.$$

For a function ϕ we define ϕ^+ by

$$\phi^+(t) = (\phi(t))^+.$$

It follows from the above that if $\{g_n\}$ is a Cauchy sequence of elements of $\mathcal{C}([0,1])$ then so is g_n^+. So if $g_n \to g$ in $\mathbf{H_R}$, then g_n^+ converges to some element, call it g^+. We want to show that $g^+ = g$. Since $g_n^+ - g_n \geq 0$, we have

$$(g, g_n^+ - g_n) \geq 0$$

by condition 2, and passing to the limit we get

$$\|g\|_2^2 = (g, g) \leq (g, g^+).$$

By the Cauchy-Schwarz inequality the right hand side is $\leq \|g\|_2 \|g^+\|_2$ so we get

$$\|g\|_2 \leq \|g^+\|_2.$$

On the other hand, from its very definition we know that

$$\|g_n^+\|_2^2 \leq \|g_n\|_2^2$$

since $|g_n^+(t)| \leq |g_n(t)|$ for all t. Passing to the limit gives $\|g^+|_2 \leq \|g\|_2$ so

$$\|g\|_2 = \|g^+\|_2.$$

But then

$$\|g-g^+\|_2^2 = \|g\|_2^2+\|g^+\|_2^2-2(g,g^+) \le |g^+\|_2^2+\|g\|_2^2-2\|g\|_2^2 = \|g^+\|_2^2-\|g\|_2^2 = 0$$

so $g = g^+$.

In short, it makes sense to ask whether or not $g \succeq 0$ for any $g \in \mathbf{H_R}$. That is, it makes sense to talk of non-negative random variables. If $g_n \in \mathbf{H_R}$ are non-negative random variables, and $g_n \to g$ then it follows from the first definition above that $g \ge 0$. In other words, the set of non-negative random variables is a closed subset of $\mathbf{H_R}$. The sum of two non-negative random variables is again non-negative, and the product of a non-negative random variable by a non-negative real number is again non-negative. Applying the first definition to f and the second definition to g we see that if $f \succeq 0$ and $g \succeq 0$ then $(f,g) \ge 0$.

In fact, the above proof shows that for *any* $g \in \mathbf{H_R}$ we can define its "non-negative part" g^+ as the limit of g_n^+ where $g_n \in \mathcal{C}([0,1]) \to g$ and define its non-positive part g^- as $(-g)^+$ and its absolute value $|g|$ as $|g| := g^+ + g^-$. Equally well we could define $|g|$ to be the limit of $|g_n|$ if g_n is a sequence of elements in $\mathcal{C}[0,1]$ converging to g. We have $\|g^+\|_2 \le \|g\|_2$ and $\|g^-\|_2 \le \|g\|_2$ while $|\ |g\ \|_2 = \|g\|_2$. We can also check that the map $g \mapsto g^+$ is continuous, and hence so is the map $g \mapsto |g|$.

We can now write

$$f \succeq g \quad \text{iff } f-g \succeq 0$$

and so

$$f \succeq g \ \Rightarrow f+h \succeq g+h$$

for any $h \in \mathbf{H_R}$.

For any real numbers a and b we have

$$\max\{a,b\} = \frac{1}{2}(a+b+|a-b|) \ \text{ and } \quad \min\{a,b\} = \frac{1}{2}(a+b-|a-b|).$$

So if f and g are functions, and we define

$$f \vee g := \frac{1}{2}(f+g+|f-g|), \quad f \wedge g := \frac{1}{2}(f+g-|f-g|),$$

then $(f \vee g)(t) = \max\{f(t), g(t)\}$ and $(f \wedge g)(t) = \min\{f(t), g(t)\}$. But the above defintions make sense in $\mathbf{H_R}$ and we shall use them.

For example, if f and g are elements of $\mathbf{H_R}$ and $f \succeq 0$ and $g \succeq 0$, then we can find $f_n \succeq 0$, $\ g_n \succeq 0$ in $\mathcal{C}([0,1])$ with $f_n \to f$ and $g_n \to g$. We know that $(f_n \wedge g_n)(t) \ge 0$ for all t, and so $f_n \wedge g_n \succeq 0$. From the continuity of the map $u \mapsto |u|$ we conclude that $f \wedge g \succeq 0$.

Therefore, if $f \succeq 0$ and $g \succeq 0$ then

$$(f,g) \ge (f, f\wedge g) \ge (f\wedge g, f\wedge g) > 0 \ \text{ unless } f \wedge g = 0.$$

Also, $f \vee g$ for any $f, g \in \mathbf{H_R}$ can be characterized as being the "least upper bound" of f and g in the sense that $f \vee g \succeq f$, $f \vee g \succeq g$, and if $h \succeq f$ and $h \succeq g$

then $h \succeq (f \vee g)$. Similarly $f \wedge g$ can be characterized as being the "greatest lower bound" of f and g.

If $f, g, h \in \mathbf{H_R}$ then

$$\begin{aligned} f - (g \vee h) &= f - \frac{1}{2}(g + h + |g - h|) \\ &= \frac{1}{2}(f - g + f - h - |f - g - (f - h)|) \\ &= (f - g) \wedge (f - h). \end{aligned} \tag{14}$$

Similarly, if follows directly from the definition that the "distributive laws"

$$f \wedge (g \vee h) = (f \wedge g) \vee (f \wedge h) \tag{15}$$

$$f \vee (g \wedge h) = (f \vee g) \wedge (f \vee h) \tag{16}$$

hold.

Finally, from the "triangle inequality" $|u + v| \preceq |u| + |v|$ one can conlude that the oprations $f \vee g$ and $f \wedge g$ are continuous, e.g. $f_n \to f$ and $g_n \to g$ implies that $f_n \vee g_n \to f \vee g$ and similarly of $\wedge$. We could keep going on, so I will just mention that from the definitions we have

$$f + g = f \wedge g + f \vee g. \tag{17}$$

8. Put in the details for these and similar arguments until you get tired. But be sure to include the preceding equation and the following consequence both of which we will use:

$$\text{If } f \wedge g = 0 \text{ then } f \vee g = f + g. \tag{18}$$

4.1 The Boolean distributive laws.

In set theory we have the following "distributive law" for unions and intersections of sets:

$$S \cap \left(\bigcup_\alpha S_\alpha \right) = \bigcup_\alpha (S \cap S_\alpha)$$

with a similar distributive law with $\cap$ and $\cup$ interchanged.

We wish to show that similar formulas hold for $\vee$ and $\wedge$. We have characterized $f \wedge g$ as the least upper bound of f and g. More generally, suppose we are given a collection of elements $x_\alpha \in \mathbf{H_R}$. We will say that a is the least upper bound of this set and write

$$a = \bigvee_\alpha x_\alpha$$

if

- $a \succeq x_\alpha$ for all α and

- If $b \succeq x_\alpha$ for all α then $b \succeq a$.

If there are finitely many x_α then we know that a exists. We will need to deal with the case of infinitely many x_α where such an a may or may not exist. Let us assume that a exists. Then for any $x \in \mathbf{H_R}$, $x_\alpha \preceq a$ implies that $x \wedge x_\alpha \preceq x \wedge a$, so $x \wedge a$ is an upper bound for all the $x \wedge x_\alpha$. We want to show that it is a (and hence the) least upper bound. For this we will use the fact (17):

$$x + x_\alpha = x \wedge x_\alpha + x \vee x_\alpha$$

in the form

$$x \wedge x_\alpha = x + x_\alpha - (x \vee x_\alpha).$$

Thus if $y \succeq x \wedge x_\alpha$ then

$$x + x_\alpha - x \vee x_\alpha \preceq y$$

or

$$x_\alpha \leq y - x + (x \vee x_\alpha).$$

Hence

$$a \preceq y - x + x \vee a.$$

9. Proceeding to argue this way, show that $x \wedge a$ is indeed the least upper bound of the x_α.

In symbols, we can write this as

$$x \wedge \left(\bigvee_\alpha x_\alpha \right) = \bigvee_\alpha (x \wedge x_\alpha) \tag{19}$$

in the sense that the right hand side exists and is equal to the left hand side under the assumption that the left hand side exists.

4.2 More Events.

We shall call a random variable $e \in \mathbf{H_R}$ an **event** if $0 \preceq e \preceq \mathbf{1}$ and

$$e \wedge (\mathbf{1} - e) = 0.$$

Then

$$e \vee (\mathbf{1} - e) = -(-e \wedge (-\mathbf{1} + e)) = \mathbf{1} - \mathbf{1} - (-e \wedge (-\mathbf{1} + e) =$$

$$= \mathbf{1} - (\mathbf{1} - (-e \wedge (-\mathbf{1} + e))) = \mathbf{1} - ((\mathbf{1} - e) \wedge e) = \mathbf{1}.$$

Conversely, if

$$e \vee (\mathbf{1} - e) = \mathbf{1}$$

then

$$e \wedge (\mathbf{1} - e) = 0$$

by reversing the argument.

Suppose that $\{e_\alpha\}$ is a family of events, and suppose that this family has a least upper bound, i.e. suppose there is an $e \in \mathbf{H_R}$ with the property that

$$e \succeq v_\alpha, \ \forall \alpha$$

and

$$y \succeq v_\alpha \ \forall \alpha \ \Rightarrow y \succeq e.$$

The usual notation is to write

$$e = \bigvee_\alpha e_\alpha$$

if such a least upper bound exists. We claim that e, if it exists, is an event. Indeed, since all the $e_\alpha \succeq 0$ we see that $e \succeq 0$, and since all the $e_\alpha \preceq \mathbf{1}$ we see that $e \preceq \mathbf{1}$. So

$$0 \preceq \mathbf{1} - e \preceq \mathbf{1} - e_\alpha$$

so from

$$e_\alpha \wedge (\mathbf{1} - e_\alpha) = 0$$

we conclude that

$$e_\alpha \wedge (\mathbf{1} - e) = 0$$

for all α. Therefore by the distributive law,

$$e \wedge (\mathbf{1} - v) = \left(\bigvee_\alpha e_\alpha \right) \wedge (\mathbf{1} - e) = \bigvee_\alpha (e_\alpha \wedge (\mathbf{1} - e)) = 0.$$

All this was quite general, and left open the issue of whether the least upper bound exists. We know it exists for a finite family. We will now use the Hilbert space structure to show that it exists for a countable family of events.

This is the replacement (via Hilbert space) of the countable additivity of measures that we will been doing in class in a couple of weeks.

So suppose that $e_i, \ i = 1, 2, 3 \ldots$ is a sequence of events.

10. Show that the

$$f_n := \bigvee_1^n e_n$$

form a Cauchy sequence, and then that their limit is the least upper bound of the $\{e_i\}$ and so is an event. [Hint: Begin by showing that $0 \preceq f_n - f_m \preceq \mathbf{1}$ for $n > m$ and then comparing $\|f_n - f_m\|^2$ with $(f_n, \mathbf{1}) - (f_m, \mathbf{1})$.]

Here is another illustration of the interplay between the Hilbert space structure and the order structure.

Suppose that $\mathbf{C}$ is a closed subset of $\mathbf{H_R}$ which has an upper bound x. In other words,

$$x \succeq u \ \ \forall\, u \in \mathbf{C}.$$

Suppose also that $\mathbf{C}$ is closed under the $\vee$ operation. In other words, assume that

$$y \in \mathbf{C}, z \in \mathbf{C} \ \Rightarrow y \vee z \in \mathbf{C}.$$

Then,

11. Under the above hypotheses, show that the least upper bound $\bigvee \mathbf{C}$ exists and belongs to $\mathbf{C}$. [Hint: Since $\mathbf{C}$ is closed, we can find a $z \in \mathbf{C}$ which is closest to x in the sense that

$$\|x - z\| \leq \|x - y\| \quad \forall y \in \mathbf{C}.$$

Show that z is the desired $\bigvee \mathbf{C}$ by showing that $z \vee y = z$ for any $y \in \mathbf{C}$. Show that $0 \preceq x - (y \vee z) \preceq x - z$. Use this to get $\|x - (z \vee y)\| \leq \|x - z\|$, and then expand out $\|x - z\|^2 = \|x - (y \vee z) + (y \vee z) - z\|^2$ and use the non-negativity of the cross terms.]

Consider any element $X \in \mathbf{H_R}$. Let $\mathcal{C}_X$ denote the set of all events e such that

$$e \preceq X^+.$$

This satisfies the conditions above, and hence there exists a least upper bound, which is an event. We shall denote this event by

$$\mathbf{1}(X \geq 1).$$

Similarly, for any positive number a we can consider the event $\mathbf{1}_{\frac{1}{a}X \geq 1}$ which we can write as

$$\mathbf{1}_{X \geq a}$$

and in a similar fashion we can construct events such as

$$\mathbf{1}_{a \leq X \leq b}$$

for any random variable X.

5 Conditional expectation is orthogonal projection.

This leads to the following idea: Let us call a closed linear subspace $S \subset \mathbf{H_R}$ a **probabilistically closed subspace** if

- $\mathbf{1} \in S$ and
- If X and $Y \in S$ implies that $X \wedge Y$ and $X \vee Y$ are in S.

The simplest example of a probabilistically closed subspace is the one dimensional space spanned by $\mathbf{1}$. If $Z \in \mathbf{H_R}$ is any random variable, orthogonal projection of Z onto this subspace is given by

$$Z \mapsto E(Z)\mathbf{1}.$$

Here is a slightly less simple but still elementary example: Let $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ a collection of events such that

$$\sum_1^n \mathbf{1}_{A_i} = \mathbf{1}.$$

In other words we have a finite set of mutually exclusive and exhaustive events. Then the space S that they span is the set of "step functions"

$$\sum a_i \mathbf{1}_{A_i}.$$

Notice that if

$$X = \sum a_i \mathbf{1}_{A_i} \quad \text{and} \quad Y = \sum_i b_i \mathbf{1}_{A_i}$$

then

$$X \vee Y = \sum c_i \mathbf{1}_{A_i} \quad \text{where } c_i = \max(a_i, b_i)$$

and so belongs to S, similarly for $X \wedge Y$. Orthogonal projection onto the space S is given by

$$\pi_S : Z \mapsto \sum_{i=1}^n \frac{(Z, \mathbf{1}_{A_i})}{P(A_i)} \mathbf{1}_{A_i}.$$

The rest of these notes/ problem set will only become clear as we progress in the course. In terms of some language that we will introduce later, the image $\pi(Z)$ is "measurable" with respect to the (finite) σ-field of events generated by the events $\mathbf{1}_{A_i}$. The finite valued element $\pi(Z)$ is called the conditional expectation of Z relative to this finite σ-field.

So in general, given any probabilistically closed subspace S, we call orthogonal projection onto S the **conditional expectation** relative to S. It is sometimes written as

$$\pi_S(Z) = E(Z \| S).$$

In general, suppose we are give a family of probabilistically closed subspaces. Then their intersection is again a probabilistically closed subspace. So given any subset B of $\mathbf{H_R}$, the intersection of all probabilistically closed subspaces contain it will be called the **probabilistically closed subspace generated** by B. For instance, if X is any element of $\mathbf{H_R}$, the probabilistically closed subspace generated by X will contain all the events $\mathbf{1}_{\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}}$. This is a rather subtle notion.

For instance, suppose we believe in MATLAB's capability of producing $(X, Y) = \text{rand}(1,2)$. (It is easy enough to see that we can theoretically produce such a pair of uniformly distributed random variables over the unit square

in the plane.) Then the events in the probabilistically closed subspace generated by X, model all subsets of the unit square of the form $A \times I$ where A is a subset of the unit interval; in other words sets of the form $\{(x,y)|x \in A\}$.

We can now also write down the general notion of **independence** of two random variable in terms of these concepts: Let S_X be the probabilistically closed subspace generated by X, and let S_Y be the probabilistically closed subspace generated by Y. Then X and Y are independent if and only if

$$\text{cov } (f,g) = 0 \quad \forall f \in S_X, \quad g \in S_y.$$

Notice that this independence a much more stringent condition than being uncorrelated which merely says that cov $(X,Y) = 0$.

My point is that the concept of orthogonal projection is a vast generalization of (an important special case of) some very subtle concepts in probability theory.

6 Simpson's "paradox".

The following (the rest of these notes) contains no problems and does not involve any serious mathematics (other that properties of fractions). But I am including it as part of your "general education" as a citizen.

6.1 Statement and example.

Let A, B, C be events. It is possible that

$$\begin{aligned} P(A|B \cap C) &< P(B^c \cap C) \\ P(A|B \cap C^c) &< P(B^c \cap C^c) \\ \text{yet} & \\ P(A|B) &> P(A|B^c). \end{aligned}$$

Here is an example which approximates an anti-discrimination case against the University of California at Berkeley:

Assume that there are two departments, History and Geography, with A denoting acceptance of applicants. Let B denote the applicant being male so B^c denotes the applicant being female. Let C denote the application is to the History department and C^c that the application is to the geography depatment (assuming that these are mutually exclusive and exhaustive).

Assume that there were 5 male applicants to the History department of which one was accepted and 8 female applicants of which 2 were accepted. (If these numbers look unrealistically small, multiply all of them by 10 or 100.)

For the Geography department there were 8 male applicants of whom 6 were accepted and 5 female applicants of whom 4 were accepted. So

$$\begin{aligned} P(A|B \cap C) &= 1/5 \\ P(A|B^c \cap C) &= 2/8 \\ P(A|B \cap C^c) &= 6/8 \\ P(A|B^c \cap C^c) &= 4/5. \end{aligned}$$

So the acceptance rate for females is higher in each department than for males. Yet overall, the total number of male and of female applicants is 13, and 7 of the thirteen male applicants is accepted while only 6 of the thirteen female applicants is accepted. So $P(A|B) > P(A|B^c)$.

The assertion about fractions is

$$1/5 < 2/8, \quad 6/8 < 4/5 \quad \text{but } 7/13 \ > \ 6/13.$$

In other words we can have eight positive numbers a, b, c, d, A, B, C, D with

$$\frac{a}{b} < \frac{A}{B}, \quad \frac{c}{d} < \frac{C}{D} \ \text{ yet } \ \frac{a+b}{c+d} > \frac{A+B}{C+D}.$$

Put still another way, let

$$u = (\,b\,)\,a, \ v = (\,d\,)\,c, \ U = (\,B\,)\,A, \ V = (\,D\,)\,C$$

be vectors in the plane (in the first quadrant). We can have

$$\text{slope } u < \text{slope } U, \ \text{slope } v < \text{slope } V$$

yet

$$\text{slope } (u+v) > \text{slope } (U+V).$$

6.2 A real life example.

From Cohen and Nagel, *An introduction to logic and scientific method* (1934). In 1910

- The death rate from tuberculosis among African Americans was *lower* in Richmond Virginia than in New York City.
- The death rate from tuberculosis among non-African Americans was lower in Richmond than in New York City, yet
- The overall death rate from tuberculosis in New York City was lower than in Richmond.

6.3 The danger of misapplication.

The comes from a misapplication of the dubious notion of "causation". Of thinking of a conditional probability as indicating a "cause", and/or of thinking of $P(A|B)$ as being a probability measure of B instead of applying Bayes to get $P(B|A)$.

6.4 Savage's "sure thing" principle.

The great philosopher of probability, L. J. Savage, in his book *The foundations of statistics* (1954) (Wiley) pp. 21-22 formulates the **sure thing principle** as follows:

> If you would definitely prefer g to f, either knowing that the event C obtained or knowing that the event C did not obtain, then you definitely prefer g to f.

As we have seen, if $g = P(A|B^c)$ and $f = P(A|B)$ and "prefer" means $g > f$ then the "sure thing principle" need not be true.

In fact, the name "Simpson's paradox" is taken from the paper Blyth, C.R. 1972 "On Simpson's paradox and the sure thing principle" *Journal of the American Statistical Association* **67** 364-366. But the effect was understood much earlier, in fact by Yule and others around 1900.

6.5 More explanation.

We have (assuming all denominators are positive)

$$P(A|B) = \frac{P(A \cap B)}{P(B} = \frac{P(A \cap B \cap C) + P(A \cap B \cap C^c)}{P(B)}$$

$$= \frac{P(A \cap B \cap C)}{P(B \cap C)} \frac{P(B \cap C)}{P(B)} + \frac{P(A \cap B \cap C^c)}{P(B \cap C^c)} \frac{P(B \cap C^c)}{P(B)}$$

so

$$P(A|B) = P(A|B \cap C)P(C|B) + P(A|B \cap C^c)P(C^c|B)$$

and similarly,

$$P(A|B^c) = P(A|B^c \cap C)P(C|B^c) + P(A|B^c \cap C^c)P(C^c|B^c).$$

Suppose that

$$P(A|B \cap C) < P(A|B^c \cap C) < P(A|B \cap C^c) < P(A|B^c \cap C^c)$$

and that $P(C|B)$ is close to zero while $P(C|B^c)$ is close to one.

In our Berkeley example, suppose that males are intrinsically more likely to apply to the geography department that to the history department and vice versa for females. (I realize that this is a dangerous hypothesis which might cost me my job - like Summers). This would explain the effect.