STA 447 Lecture Notes

Mark Koudstaal

January 27, 2011

1 1/25/2011

1.1 Tail Events, Kolmogorov 0-1

- If we have any collection of σ-algebras $\{\mathcal{G}_n\}_{n>0}$ (where $\mathcal{G}_n \subset \mathcal{F}$ and $(\Omega, \mathcal{F}, P)$ is our probability space) then we may form the **tail σ-algebra** of this collection as follows:
  - we set $\mathcal{T}_n = \sigma\left(\bigcup_{k\geq n} \mathcal{G}_k\right)$
  - we define the **tail σ-algebra** of the collection $\{\mathcal{G}_n\}_{n>0}$, $\mathcal{T}$, as:

$$\mathcal{T} = \bigcap_{n>0} \mathcal{T}_n$$

- Now, by construction, we have $\mathcal{T}_{j+1} \subset \mathcal{T}_j$, $\forall j$ and so if $A \in \mathcal{T}$ this says that $\forall N$, $A \in \mathcal{T}_n$ (and so, effectively, all of the $\mathcal{G}_n$) for $n \geq N$ and that this characterizes A. So A is effectively determined by the aymptotic structure of the $\mathcal{G}_n$; It must keep occurring past every n.

- If it happens that this collection of σ-algebras $\{\mathcal{G}_n\}_{n>0}$, is **independent**, in the sense that for any p and distinct $n_1, \ldots, n_p \in \mathbb{N}$, if $A_{n_1} \in \mathcal{G}_{n_1}, \ldots, A_{n_p} \in \mathcal{G}_{n_p}$ then we have:

$$P\left(\bigcap_{i=1}^{p} A_{n_i}\right) = \prod_{i=1}^{p} P(A_{n_i})$$

Then we get a great result about the tail σ-algebra, $\mathcal{T}$, generated by the $\{\mathcal{G}_n\}_{n>0}$. This fabulous result was handed down to us by the godfather of modern probability: Kolmogorov. It says that if this collection of σ-algebras is independent then the events in the tail algebra effectively either happen, or do not!

So, formally, here it is:

Theorem 1. Kolmogorov's 0-1 Law: *If $\{\mathcal{G}_n\}_{n>0}$ is a collection of independent σ-algebras (on $(\Omega, \mathcal{F}, P)$) and $\mathcal{T}$ is the tail σ-algebra which they form then any $A \in \mathcal{T}$ satisfies $P(A) = 0$ or 1.*

There are many extensions of this (can even extend it to tail σ-algebra of a "nice" class of markov chains (which you'll see later)). The proof of this result follows a few simple steps:

  - For each n consider $\sigma\left(\bigcup_{k\leq n-1} \mathcal{G}_k\right)$ and $\mathcal{T}_n$.
  - The collection of events $\{A_n A_{n+1} \ldots A_{n+r} : A_k \in \mathcal{G}_k, r \geq 0\}$ generates $\mathcal{T}_n$ while $\{A_1 A_2 \ldots A_{n-1} : A_k \in \mathcal{G}_k\}$ generates $\mathcal{H}_n := \sigma\left(\bigcup_{k\leq n-1} \mathcal{G}_k\right)$.
  - The definition of independence, combined with the fact that these are π-systems, guarantees that $\mathcal{H}_n$ is independent of $\mathcal{T}_n$ for every n.
  - We have that $\mathcal{T} \subset \mathcal{T}_n$ for all n (by definition) and $\mathcal{H}_n \subset \mathcal{H}_{n+1}$ implies that $\bigcup_k \mathcal{H}_k$ is a π-system which is independent of $\mathcal{T}$ and generates $\mathcal{H}_\infty := \sigma\left(\bigcup_{k>0} \mathcal{G}_k\right)$. As the

$\mathcal{H}_n$ are independent of the $\mathcal{T}_n$ for every n, this guarantees that $\mathcal{H}_\infty$ is independent of $\mathcal{T}$.

- This relies on the following theorm: If $\mathcal{I}$ is a π-system (meaning $I_1, I_2 \in \mathcal{I} \Rightarrow I_1 I_2 \in \mathcal{I}$) and $\mathcal{F} = \sigma(\mathcal{I})$ then if $\mu_1 = \mu_2$ on $\mathcal{I}$ it follows that $\mu_1 = \mu_2$ on $\mathcal{F}$. This holds for other properties, such as independence (but this is really just an extension of this theorem: i.e. $\mu_1(A_1 A_2 \dots A_k) = P(A_1 A_2 \dots A_k)$ and $\mu_2(A_1 A_2 \dots A_k) = P(A_1)P(A_2) \cdots P(A_k)$). Really just a practical theorem that allows us to prove things about measures using tangible classes of sets.
- But $\mathcal{H}_\infty := \sigma\left(\bigcup_{k>0} \mathcal{G}_k\right)$ and so, by construction, $\mathcal{T} \subset \mathcal{H}_\infty$. The previous results thus imply that $\mathcal{T}$ is independent of itself! So for any $A \in \mathcal{T}$ we have:

$$P(AA) = P(A)P(A) \Rightarrow P(A) = [P(A)]^2$$

  i.e. we must have $P(A) = 0$ or 1.
- as a side note, this can be proven using martingale convergence theorems, which you will cover later in the course!

- Despite all of this abstraction, the tail σ-algebra is a very useful construction. Indeed, it contains many events which are useful to the study of sums of random variables. E.g. can you show that the following events are members of the tail σ-algebra of a collection of random variables $X_1, X_2, \dots$ (define this for them):
  - $\{\omega \in \Omega : \lim_n X_n(\omega) \text{ exists}\}$
  - $\{\omega \in \Omega : \sum_n X_n(\omega) \text{ converges}\}$
  - $\{\omega \in \Omega : \lim_k k^{-1} \sum_{n \leq k} X_n(\omega) \text{ exists}\}$

Kolmogorov 0-1 tells us that if the X_i are independent, then these events either happen, or do not! Similarly, there are random variables, such as:

$$Z = \limsup_n \left\{ k^{-1} \sum_{m \leq k} X_m \right\}$$

which are $\mathcal{T}$ measurable. This can prove to be very useful. E.g. can you show that the kolmogorov 0-1 law implies that if the X_i are independent then Z (or any r.v. which is $\mathcal{T}$ measurable) must be a.s. constant (possibly infinite)?

- **Example:** Suppose $X_1, X_2, \ldots$ are IID $X \sim$exp(1). Then it follows that:

$$P\left(\limsup_n \frac{X_n}{\log n} = 1 \right) = 1$$

to see this, notice that it is sufficient to show that for each $\epsilon > 0$

$$P(X_n > (1-\epsilon)\log n, i.o.) = 1 \text{ while } P(X_n > (1+\epsilon)\log n, i.o.) = 0$$

For the first statement we can employ Borel Cantelli II. Notice that:

$$P(X_n > (1-\epsilon)\log n) = P(X > (1-\epsilon)\log n) = e^{-(1-\epsilon)\log n} = n^{-(1-\epsilon)}$$

and $\sum_n n^{-(1-\epsilon)} = \infty$ for each $\epsilon > 0$. BCII thus implies that $P(X_n > (1-\epsilon)\log n, i.o.) = 1$. For the second statement, BCI would be easiest but we can use K0-1. Notice that, by definition, for all N we have:

$$\begin{aligned} P(X_n > (1+\epsilon)\log n, i.o.) &\leq \sum_{n \geq N+1} P(X_n > (1+\epsilon)\log n) \\ &= \sum_{n \geq N+1} n^{-(1+\epsilon)} \leq \int_N^\infty x^{-(1+\epsilon)} dx = \frac{1}{\epsilon N^\epsilon} \end{aligned}$$

Choosing N large enough ($N > (1/\epsilon)^{(1/\epsilon)}$) we see $P(X_n > (1+\epsilon)\log n, i.o.) < 1$. As this is a tail event, and the X_n are independent, K0-1 then implies that $P(X_n > (1+\epsilon)\log n, i.o.) = 0$.

Probably easier to just apply Borel Cantelli lemmas for computations, but K0-1 tells us that many important r.v.'s and their probabilities are trivial!

- Notice what K0-1has to say about Borel Cantelli: For a sequence of independent events A_n, set $\mathcal{G}_n = \{\emptyset, A_n, A_n^c, \Omega\}$ which are then independent σ-algebras. K0-1 tells us that, as $\{A_n, i.o.\}$ is a tail event w.r.t. this sequence of σ algebras, $P(A_n, i.o.) = 0$ or 1. The Borel Cantelli lemmas tell us precisely when this happens.

1.2 Some Integration Results:

Good bit of work to develop these from scratch. We'll just state them:

Theorem 2. Monotone Convergence: *If* $0 \leq X_n \nearrow X$ *a.s., then* $E(X) = E(\lim X_n) = \lim E(X_n)$.

Theorem 3. Fatou's Lemma: *If* $X_n \geq 0$ *a.s. then* $E(\lim_n \inf X_n) \leq \lim_n \inf E(X_n)$.

See lecture notes for a quick and easy proof of Fatou. Can use monotone convergence if you set things up properly.

Proof. If we set $Z_N = \inf_{k \geq N} X_n$ then the Z_n form an increasing sequence. Monotone convergence thus implies that:

$$E(\lim \inf X_n) = E(\lim Z_n) = \lim E(Z_n)$$

But, by construction $Z_N = \inf_{k \geq N} X_n \leq X_m$ for all $m \geq N$ and so monotonicity of integral implies that for $m \geq N$:

$$0 \leq E(Z_N) \leq E(X_m)$$

and hence:

$$0 \leq E(Z_N) \leq \inf_{m \geq N} E(X_m)$$

Taking limits gives the result. □

Theorem 4. Dominated Convergence: *If* $X_n \to X$ *a.s. and* $|X_n| \leq Y$ *with* $E(Y) < \infty$, *then* $E(X) = E(\lim X_n) = \lim E(X_n)$.

Theorem 5. Probabilistic Dominated Convergence: *If* $X_n \xrightarrow{P} X$ *and* $|X_n| \leq Y$ *with* $E(Y) < \infty$, *then* $E(X) = E(\lim X_n) = \lim E(X_n)$.

1.3 L^2 Theory and Conditional Expectation

- Work on a probability space $(\Omega, \mathcal{F}, P)$ and define $L^2(\Omega, \mathcal{F}, P)$ to be collection of random variables $X \in m\mathcal{F}$ which satisfy $EX^2 < \infty$. For such X define $||X|| = (EX^2)^{1/2}$, then this defines a pseudo norm on $L^2(\Omega, \mathcal{F}, P)$ (just quotient things out to get a norm).

- Can show that for this collection $< X, Y > := E(XY)$ defines an inner product which gives rise to $||\cdot||$ (check this). Given completeness, we have that $L^2(\Omega, \mathcal{F}, P)$ with this inner product is a Hilbert Space.

- All seems very abstract, but this lends a nice geometric flavor to the whole theory as closed subspaces and convex sets now have a closest point to any given point: their projection and this is well defined and unique. Maybe give example of why you need to be careful in infinite dimensional spaces (and possible example from Cheney about spaces in which for such sets, no closest point exists).

- So here we have it:

Theorem 6. Completeness of $L^2(\Omega, \mathcal{F}, P)$: *For each sequence of variables* $X_n \in L^2(\Omega, \mathcal{F}, P)$ *which satisfy:*

$$||X_n - X_m|| \to 0 \text{ as } m, n \to \infty$$

there is an $X \in L^2(\Omega, \mathcal{F}, P)$ *such that* $||X_n - X|| \to 0$ *as* $n \to \infty$*, and hence* $L^2(\Omega, \mathcal{F}, P)$ *is complete (and thus a Hilbert Space).*

Proof. The idea is to pick a subsequence which converges for a.e. ω and then to show that the entire sequence must converge to this value.

- So use the cauchy property to pick a subsequence for which:

  $$||X_{n+1} - X_n|| < 2^{-n}$$

  and define:

  $$Y(\omega) = |X_1(\omega)| + \sum_{n=1}^{\infty} |X_{n+1} - X_n|$$

  Then, by construction, monotone convergence implies that (Take $Y_n = |X_1| + \sum_{j=1}^{n-1} |X_{j+1} - X_j| \nearrow Y$)

  $$||Y|| \leq ||X_1|| + \sum_{n=1}^{\infty} ||X_{n+1} - X_n|| < \infty$$

  which implies that Y^2 (and hence Y) is finite, and hence converges, for a.e. ω.

- Set $X(\omega) = \lim_n X_n(\omega) = X_1(\omega) + \sum_{n=1}^{\infty}(X_{n+1}(\omega) - X_n(\omega))$. As $|X_n| \leq Y$ and Y is integrable (Cauchy Schwarz), dominated convergence implies that X is integrable and hence converges for a.e. ω. For all n, $||X_n|| \leq ||Y||$ and so another application of dominated convergence yields $||X|| \leq ||Y||$ and hence $X \in L^2(\Omega, \mathcal{F}, P)$

– Now notice that the triangle inequality implies that (taking m to be any number in the sequence and $n(m)$ to be the closest number along our specially chosen subsequence):

$$\begin{aligned} ||X_m - X|| &\leq ||X_m - X_{n(m)}|| + ||X_{n(m)} - X|| \\ &\leq ||X_m - X_{n(m)}|| + \sum_{k>n(m)} 2^{-k} \end{aligned}$$

which can be made arbitrarily small by making m large enough.

– Thus we have constructed an $X \in L^2(\Omega, \mathcal{F}, P)$ so that $||X_m - X|| \to 0$ which shows that this space is complete.

□

- One can now use this completeness to prove that nice euclideanesque properties hold for $L^2(\Omega, \mathcal{F}, P)$. Chief among these is the **closest point property** which we can prove using two key characteristics of Hilbert Space: 1.) the fact that a hilbert space has an inner product (and hence the parallelogram law holds), 2.) the fact that a hilbert space is complete:

- So here it is:

Theorem 7. Closest point property *Suppose that M is a closed linear subspace of $L^2(\Omega, \mathcal{F}, P)$. Then for each $X \in L^2(\Omega, \mathcal{F}, P)$ there is a unique $P_M X \in L^2(\Omega, \mathcal{F}, P)$ so that:*

$$||X - P_M X|| = \inf_{Z \in M} ||X - Z||$$

Furthermore $< Z, X - P_M X >= 0$ *for every* $Z \in M$. *This means that every* $X \in L^2(\Omega, \mathcal{F}, P)$ *has a unique decomposition as* $X = P_M X + (X - P_M X)$ *(and thus* $L^2(\Omega, \mathcal{F}, P)$ *has a well defined orthogonal decomposition).*

Maybe give example to show how crucial closed property is.... span of a basis example...

Proof. The idea of the proof is to pick a sequence in M, $\{X_n\}$, for which $||X_n - X||$ converges to $d = \inf_{Z \in M} ||X - Z||$ (which we may always do, by def of infimum). We then use the parallelogram law to show that this sequence must be Cauchy, and Completeness + closedness to show that it then converges to an element of $L^2(\Omega, \mathcal{F}, P)$. So here goes:

- Set $d = \inf_{Z \in M} ||X - Z||$. We may employ parallelogram law to see that:

$$\begin{aligned} ||(X - X_m) - (X - X_n)||^2 \quad &+ \quad ||2X - (X_n + X_m)||^2 \\ &= \quad 2(||X - X_m||^2 + ||X - X_n||^2) \end{aligned}$$

  Using that $||2X - (X_n + X_m)||^2 \geq 4d^2$ we arrive at:

$$||X_n - X_m||^2 + d^2 \leq 2(||X - X_m||^2 + ||X - X_n||^2)$$

  from which we see that $||X_n - X_m||$ may be made arbitrarily small by choosing m, n large enough. Thus the sequence X_n is cauchy and hence, by completeness of L^2 and closedness of M converges to an element, $P_M X$, of M.

- To see uniqueness, suppose that Y is any other element which satisfies $||X_n - Y|| \to 0$. Then we have:

$$||P_M X - Y|| \leq ||P_M X - X_n|| + ||X_n - Y||$$

which can be made arbitrarily small. Since $||X|| = 0$ implies $X = 0$, this gives $Y = P_M X$.

- To show that $< Z, X - P_M X >= 0$ for every $Z \in M$. Trivially satisfied for $Z = 0$ so suppose $Z \neq 0$ and suppose, for contradiction, that this is not the case. Then for each $\delta \in \mathbb{R}$ we find that (by the fact that $P_M X$ is closest to X):

$$d^2 = ||X - P_M X||^2 \leq ||X - P_M X - \delta Z||^2 = d^2 + \delta^2 ||Z||^2 - 2\delta < Z, X - P_M X >$$

and so we must have:

$$0 \leq \delta^2 ||Z||^2 - 2\delta < Z, X - P_M X >$$

by choosing δ small enough and opposite the sign of $< Z, X - P_M X >$ arrive at a contradiction and so conclude that $< Z, X - P_M X >= 0$. E.g. take $0 < |\delta| < 2 < Z, X - P_M X > /||Z||^2$

- Notice that at no point in this proof did we make reference to any property of $L^2(\Omega, \mathcal{F}, P)$ other than completeness, which is one of the defining properties of a hilbert space. This proof goes over unchanged to general Hilbert spaces!

□

- **Application:** One of the main applications of the closest point property to probability theory is to rigorously define the notion of conditional expectation. We work on $(\Omega, \mathcal{F}, P)$. If we are given a σ algebra $\mathcal{G} \subset \mathcal{F}$ then for $X \in L^1(\Omega, \mathcal{F}, P)$ we define $E(X|\mathcal{G})$ as the equivalence class $Z \in L^1(\Omega, \mathcal{G}, P)$ satisfying:

$$E(ZY) = E(XY)$$

for every $Y \in L^1(\Omega, \mathcal{G}, P)$. One can show that this is equivalent to requiring that:

$$E(Z1_A) = E(X1_A)$$

for every $A \in \mathcal{G}$ (one way is easy, the other is a relatively simple application of the std. machine).

- Notice that if everything is square integrable, this has the interpretation of $< Z, Y >=< X, Y >$ (in terms of our inner product $< X, Y >= E(XY)$) which is one of the defining properties of a projection (and which our closest point to a linear subspace in a hilbert space satisfies). Thus the $E(X|\mathcal{G})$ (as defined) loosely has the interpretation of a projection of X onto $L^1(\Omega, \mathcal{F}, P)$, (and so a closest point) which is sort of what motivates the definition in the first place.

  This definition is all well and good, but for a given X does such a Z even exist? If so, is it unique? By rephrasing the problem in terms of Hilbert Space theory, we can get quick answer. Here is how to proceed:

  - First notice that $L^2(\Omega, \mathcal{G}, P)$ is a closed subspace of $L^2(\Omega, \mathcal{F}, P)$ (which we have shown is a Hilbert Space). Why is this so? Take any sequence $Y_n \in L^2(\Omega, \mathcal{G}, P)$ with $||Y_n - Y|| \to 0$. Then Y is an a.s. limit of $\mathcal{G}$-measurable random variables and hence $\mathcal{G}$ measurable. Furthermore, $||Y|| \leq \sup_n(||Y_n|| + ||Y - Y_n||) < \infty$ and hence $L^2(\Omega, \mathcal{G}, P)$ is a closed subspace of $L^2(\Omega, \mathcal{F}, P)$.
  - Set $M = L^2(\Omega, \mathcal{G}, P)$. Then our previous work guarantees that for any $X \in L^2(\Omega, \mathcal{F}, P)$ there is a unique $P_M X \in L^2(\Omega, \mathcal{G}, P)$ so that:

$$||X - P_M X|| = \inf_{Y \in M} ||X - Y||$$

and $< Z, X - P_M X >= 0$ or $E(ZX) = E(ZP_M X)$ for every $Z \in L^2(\Omega, \mathcal{G}, P)$. Thus (recall uniqueness) $Z = P_M X$ satisfies the definition of conditional expectation.

- Cauchy Schwarz implies that $L^2(\Omega, \mathcal{F}, P) \subset L^1(\Omega, \mathcal{F}, P)$. Taken together with the fact that the continuous functions are dense in both $L^2(\Omega, \mathcal{F}, P)$ and $L^1(\Omega, \mathcal{F}, P)$, this implies that $L^2(\Omega, \mathcal{F}, P)$ is dense in $L^1(\Omega, \mathcal{F}, P)$. Thus for any $Y \in L^1(\Omega, \mathcal{F}, P)$ we may find a sequence $Y_n \in L^2(\Omega, \mathcal{F}, P)$ so that $|Y_n| \nearrow |Y|$ a.s. (consider $Y_n = Y I(|Y| \leq n)$)
- Given such a sequence, $E(Y_n|\mathcal{G}) = P_M Y_n$ is well defined as a version of conditional expectation. So consider $W = \lim_n E(Y_n|\mathcal{G})$. Then if we can show that
  - $W = \lim_n E(Y_n|\mathcal{G})$ satisfies the conditional expectation property
  - $W \in L^1(\Omega, \mathcal{G}, P)$

  we will have proven the existence of a random variable $E(Y|\mathcal{G})$ for every $Y \in L^1(\Omega, \mathcal{F}, P)$. Enough to show it for positive random variables, which works (using monotone convergence and fact that $U \geq 0$ implies $E(U|\mathcal{G}) \geq 0$ + alternative definition of conditional expectation in terms of coincidence of expectations over $\mathcal{G}$-sets).

- **Application:** Maybe talk about Karhunen-Lòeve .... Optimal projection, sparsest representation.... Useful for functional estimation.... catch 22: do you gain more from sparsity than you loose in estimating the optimal basis?? Should be the case with large enough data sets (properly done, estimates of eigenstucture are effectively consistent).

1.4 Generating Function Stuff

- If $X : \Omega \to \{0\} \cup \mathbb{N}$, we call X a counting random variable. A useful tool for studying the structure of such r.v.'s X is the **Probability Generating Funciton**, $G_X(s)$, defined by:

$$G_X(s) := E(s^X) = \sum_{k=0}^{\infty} s^k P(X = k)$$

- Jensen implies that $|G(s)| \leq E(|s|^X) \leq \sum_{k=0}^{\infty} P(X = k) = 1$ for $|s| \leq 1$ and, furthermore, is uniformly convergent for s in this range.

- If we assume that the moments of X are finite then we get that

$$|G^{(p)}(s)| \leq \sum_{k=1}^{\infty} |s|^k k(k-1)\dots(k-p+1)P(X = k) \leq \sum_{k=1}^{\infty} |s|^k k^p P(X = k) \leq E(X^k) < \infty$$

for $|s| \leq 1$ and so the derivatives of G are uniformly convergent for $|s| \leq 1$

- As the limits defining the $G^{(p)}(s)$ are continuous and the convergence is uniform for $|s| \leq 1$, we get that the derivatives are continuous (as uniform convergence preserves continuity).

- Combining these facts, we get that for any p, and $y \in (0, 1]$

$$G^{(p)}(y) = \lim_{s \nearrow y} G^{(p)}(s) = E[\lim_{s \nearrow y} s^X X(X-1)\cdots(X-p+1)]$$

and taking $y = 1$ we get:

$$G^{(p)}(1) = E(X(X-1)\cdots(X-p+1))$$

- Can use similar arguments to show that:

$$G^{(k)}(0) = k!P(X = k)$$

and so from the derivatives of G we may recover the distribution of X!

- May extend, probability generating functions to vectors of counting random variables: For $X : \Omega \to (\{0\} \cup \mathbb{N})^n$ and $s \in \mathbb{R}^n$, we define the PGF, $G_X(s)$, of X as:

$$G_X(s_1, \ldots, s_n) := E(s_1^{X_1} \cdots s_n^{X_n})$$

Analogous results hold. Good exercise to show that independence of $X_1, \ldots, X_n$ is equivalent to:

$$G_X(s_1, \ldots, s_n) = G_{X_1}(s_1) \cdots G_{X_n}(s_n)$$

- Suppose that N is a counting random variable with pgf G_N as are $X_1, X_2, \ldots$ (which are IID copies of some counting r.v. X, independent of N) with pgf $G_X(s)$. Consider the sum:

$$S = X_1 + \cdots + X_N$$

which you will see is useful in the study of branching processes. Then we have that:

$$G_S(s) = E(s^S) = E(s^{X_1 + \cdots + X_N}) = E(E(s^{X_1 + \cdots + X_N} | N)) = E(G_X(s)^N)$$

and so $G_S(s) = G_N(G_X(s))$.

1.5 References

- David Williams (1991) *Probability with Martingales*. CUP.

- Elias M. Stein and Rami Shakarchi (2005) *Real Analysis*. Princeton.

- Geoffrey Grimmett and David Stirzaker (2001) *Probability and Random Processes, Third Edition.* Oxford.

- Philip McDunnough *STA 447 Lecture Notes*