STA 711: Probability & Measure Theory
Robert L. Wolpert

6 Independence (cont)

6.5 B/C + Independence Illustration

Let $X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ and let $\{A_n\} \subset \mathcal{F}$ be events that satisfy $\mathsf{P}[A_n] \to 0$. Does it follow that $X_n := X\mathbf{1}_{A_n}$ converges almost-surely to 0?
If $\sum_n \mathsf{P}[A_n] < \infty$, then *yes*— by the Borel-Cantelli lemma,

$$\mathsf{P}[X_n \not\to 0] \le \mathsf{P}[\limsup A_n] = 0,$$

so $X_n \to 0$ *a.s.*
BUT, if $\{\mathsf{P}[A_n]\}$ is not summable, then *a.s.* convergence can fail. For example, if $\{A_n\}$ are independent and $X \equiv 1$, then

$$\mathsf{P}[X_n \not\to 0] \ge \mathsf{P}[\limsup A_n] = 1,$$

so $\mathsf{P}[X_n \to 0] = 0$. In Week 7 we will find a new sense of convergence called "convergence in probability" that is weaker than almost-sure convergence, and we'll show that $X_n \to 0$ *pr.*

6.6 Another Zero-One Law: Kolmogorov's

For any collection $\{X_n\}$ of random variables on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, define two sequences of σ-algebras ("past" and "future") by:

$$\mathcal{F}_n := \sigma\{X_i : \ i \le n\} \qquad \mathcal{T}_n := \sigma\{X_i : \ i \ge n+1\}$$

and, from them, construct the π-system $\mathcal{P}$ and "tail" σ-algebra $\mathcal{T}$ by

$$\mathcal{P} := \bigcup_{n=1}^{\infty} \mathcal{F}_n \qquad \mathcal{T} := \bigcap_{n=1}^{\infty} \mathcal{T}_n.$$

In general $\mathcal{P}$ will not be a σ-algebra, because it will not be closed under countable unions or intersections, but it is a field and hence a π-system, and generates the σ-algebra $\vee\mathcal{F}_n := \sigma(\mathcal{P}) \subseteq \mathcal{F}$.
The class $\mathcal{T}$, called the *tail* σ-field, includes those events that depend only on what happens *eventually*, regardless of what happens for the first few (or few million) $\{X_n\}$. These include such events as "$\{X_n$ converges$\}$" or "$\{\limsup X_n \le 1\}$" or, with $S_n := \sum_1^n X_j$, "$\{\frac{1}{n}S_n$ converges$\}$" or "$\{\frac{1}{n}S_n \to 0\}$," but not events like "$\{\min X_n \le c\}$."

Theorem 1 (Kolmogorov's Zero-One Law) *For independent random variables X_n, the tail σ-field $\mathcal{T}$ is "almost trivial" in the sense that every event $\Lambda \in \mathcal{T}$ has probability $\mathsf{P}[\Lambda] = 0$ or $\mathsf{P}[\Lambda] = 1$.*

Proof. Let $A \in \mathcal{P} = \cup \mathcal{F}_n$, and $\Lambda \in \mathcal{T}$. Then for some $n \in \mathbb{N}$, $A \in \mathcal{F}_n$ and $\Lambda \in \mathcal{T}_n$, so $A \perp\!\!\!\perp \Lambda$; thus $\mathcal{P}$ and $\mathcal{T}$ are independent. Since $\mathcal{P}$ is a π-system, it follows from the Basic Criterion that $\sigma(\mathcal{P})$ and $\mathcal{T}$ are also independent. But $\mathcal{F}_n \subset \mathcal{P}$ so each X_n is $\sigma(\mathcal{P})$-measurable, hence $\mathcal{T} \subset \sigma(\mathcal{P})$ and each $\Lambda \in \mathcal{T}$ must also be in $\sigma(\mathcal{P}) \perp\!\!\!\perp \mathcal{T}$. It follows that:

$$\mathsf{P}[\Lambda] = \mathsf{P}[\Lambda \cap \Lambda] = \mathsf{P}[\Lambda]\mathsf{P}[\Lambda] = \mathsf{P}[\Lambda]^2,$$

so $0 = \mathsf{P}[\Lambda]\big(1 - \mathsf{P}[\Lambda]\big)$ proving that $\mathsf{P}[\Lambda]$ must be zero or one. ∎

6.7 Product Spaces

Do independent random variables exist, with arbitrary (marginal) specified distributions? How can they be constructed? One way is to build *product probability spaces*; let's see how to do that.

Let $(\Omega_j, \mathcal{F}_j, \mathsf{P}_j)$ be a probability space for $j = 1, 2$ and set:

$$\begin{aligned}
\Omega = \Omega_1 \times \Omega_2 &:= \{(\omega_1, \omega_2) : \ \omega_j \in \Omega_j\} \\
\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 &:= \sigma\{A_1 \times A_2 : \ A_j \in \mathcal{F}_j\} \\
\mathsf{P} &:= \mathsf{P}_1 \otimes \mathsf{P}_2, \text{ the unique extension to } \mathcal{F} \text{ satisfying:} \\
\mathsf{P}(A_1 \times A_2) &= P_1(A_1) \cdot P_2(A_2) \text{ for } A_1 \in \mathcal{F}_1,\ A_2 \in \mathcal{F}_2.
\end{aligned}$$

Any random variables X_1 on $(\Omega_1, \mathcal{F}_1, \mathsf{P}_1)$ and X_2 on $(\Omega_2, \mathcal{F}_2, \mathsf{P}_2)$ can be extended to the common space $(\Omega, \mathcal{F}, \mathsf{P})$ by defining $X_1^*(\omega_1, \omega_2) := X_1(\omega_1)$ and $X_2^*(\omega_1, \omega_2) := X_2(\omega_2)$; it's easy to show that $\{X_j^*\}$ are independent and have the same marginal distributions as $\{X_j\}$. Thus, independent random variables do exist with arbitrary distributions. The same construction extends to countable families.

6.8 Fubini's Theorem

We now consider how to evaluate probabilities and integrals on product spaces.

For any $\mathcal{F}$-measurable random variable $X : \Omega_1 \times \Omega_2 \to \mathcal{S}$ ($\mathcal{S}$ would be $\mathbb{R}$, for real-valued RVs, but could also be $\mathbb{R}^n$ or any complete separable metric space), and for any $\omega_2 \in \Omega_2$, the (second) *section* of X is the $(\Omega_1, \mathcal{F}_1, \mathcal{P}_1)$ random variables $X_{\omega_2} : \Omega_1 \to \mathcal{S}$ defined by

$$X_{\omega_2}(\omega_1) := X(\omega_1, \omega_2).$$

It is not *quite* obvious, but true, that X_{ω_2} is $\mathcal{F}_1$-measurable— show this first for indicator random variables $X = \mathbf{1}_{A_1 \times A_2}$ of product sets, then extend by a π-λ argument to indicators

$X = \mathbf{1}_A$ of sets $A \in \mathcal{F}$, then to simple RVs by linearity, then to the nonnegative RVs X_+ and X_- for an arbitrary $\mathcal{F}$-measurable X by monotone limits. Similarly, the first section $X_{\omega_1}(\cdot) := X(\omega_1, \cdot)$ is $\mathcal{F}_2$-measurable for each $\omega_1 \in \Omega_1$.

Finally: Fubini's theorem gives conditions (namely, that either $X \geq 0$ or $\mathsf{E}|X| < \infty$) to guarantee that these three integrals are meaningful and equal:

$$\int_{\Omega_2} \left\{ \int_{\Omega_1} X_{\omega_2} d\mathsf{P}_1 \right\} d\mathsf{P}_2 \stackrel{?}{=} \iint_{\Omega} X \, d\mathsf{P} \stackrel{?}{=} \int_{\Omega_1} \left\{ \int_{\Omega_2} X_{\omega_1} d\mathsf{P}_2 \right\} d\mathsf{P}_1 \tag{1}$$

To prove this, first note that it's true for indicators $X = \mathbf{1}_{A_1 \times A_2}$ of the π-system of measurable rectangles $(A_1 \times A_2)$ with each $A_j \in \mathcal{F}_j$; then verify that the class $\mathcal{C}$ of events $A \in \mathcal{F}$ for which it holds for $X = \mathbf{1}_A$ is a λ-system. By Dynkin's π-λ theorem it follows that $\mathcal{F} \subset \mathcal{C}$ so (1) holds for all indicators $X = \mathbf{1}_A$ of events $A \in \mathcal{F}$, hence for all nonnegative simple functions in $\mathcal{E}_+$, and finally for all $\mathcal{F}$-measurable $X \geq 0$ by the MCT. For $X \in L_1$, apply this result separately to X_+ and X_-.

Fubini's theorem applies more generally. Each probability measure P_j may be replaced by an arbitrary σ-finite[1] measure:

Theorem 2 (Fubini) *Let $(\mathcal{X}, \mathcal{F}, \mu)$ and $(\mathcal{Y}, \mathcal{G}, \nu)$ be two σ-finite measure spaces, and let $f(x, y)$ be a real-valued measurable function on the product space $\big(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \times \mathcal{G}, \mu \otimes \nu\big)$. Then*

$$\int_{\mathcal{Y}} \left\{ \int_{\mathcal{X}} f(x,y)\mu(dx) \right\} \nu(dy) = \iint_{\mathcal{X}\times\mathcal{Y}} f(x,y)(\mu \otimes \nu)(dx\, dy) = \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} f(x,y)\nu(dy) \right\} \mu(dx)$$

if either

- *$f(x, y) \geq 0$ for $(x, y) \in \mathcal{N}^c$ for some $\mathcal{N} \in \mathcal{F} \times \mathcal{G}$ with $(\mu \otimes \nu)(\mathcal{N}) = 0$,* or
- *$f \in L_1\big(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \times \mathcal{G}, \mu \otimes \nu\big)$.*

Also,one of the measures (say, P_2) may be replaced by a measurable kernel[2] $K(\omega_1, d\omega_2)$ that is a σ-finite measure $K(\omega_1, \cdot)$ on $\mathcal{F}_2$ in its second variable for each fixed ω_1, and an $\mathcal{F}_1$-measurable function $K(\cdot, B_2)$ in its first variable for each fixed $B_2 \in \mathcal{F}_2$. Now Fubini's Theorem asserts (under positivity or L_1 conditions) the equality of integrals of X wrt the measure $\mathsf{P}(d\omega_1\, d\omega_2) = \mathsf{P}_1(d\omega_1)K(\omega_1, d\omega_2)$ to the iterated integrals

$$\int_{\Omega_2} \nu_X(d\omega_2) = \iint_{\Omega} X d\mathsf{P} = \int_{\Omega_1} \left\{ \int_{\Omega_2} X_{\omega_1}(\omega_2) K(\omega_1, d\omega_2) \right\} \mathsf{P}_1(d\omega_1)$$

for the measure on $\mathcal{F}_2$ given by $\nu_X(d\omega_2) := \int_{\Omega_1} X_{\omega_2}(\omega_1) K(\omega_1, d\omega_2) \mathsf{P}_1(d\omega_1)$.

[1] Recall that a measure μ on a measurable space $(\mathcal{X}, \mathcal{F})$ is "σ-finite" if there are countably-many sets $\{\Lambda_j\} \subset \mathcal{F}$ with $\mu(\Lambda_j) < \infty$ for each j, and $\mathcal{X} = \cup_j \Lambda_j$. Evidently any finite measure (including probability measures) is also σ-finite, but the converse is false. Lebesgue measure is σ-finite on $\mathbb{R}^n$, for example.

[2] Measurable kernels come up all the time when studying conditional distributions (as you'll see in week 9 of this course) and, in particular, Markov chains and processes.

As an easy consequence of Theorem 2 (take $(\mathcal{X},\mathcal{F},\mu)=(\Omega,\mathcal{F},\mathsf{P})$, let $(\mathcal{Y},\mathcal{G})=(\mathbb{N},2^{\mathbb{N}})$, and let $\nu(A):=\#\{A\}$ be counting measure for $A\subset\mathbb{N}$), for any sequence of random variables we may exchange summation and expectation and conclude that equality holds in

$$\mathsf{E}\left\{\sum_{n=1}^{\infty}X_n\right\}\stackrel{?}{=}\sum_{n=1}^{\infty}\{\mathsf{E}X_n\}$$

whenever each $X_n\ge 0$ or when $\sum_{n=1}^{\infty}\mathsf{E}|X_n|<\infty$, but otherwise equality may fail.

6.8.1 A Counter-example

For an example where interchanging integration order fails, integrate by parts to verify:

$$\int_0^1\left\{\int_0^1\frac{y^2-x^2}{(x^2+y^2)^2}\,dx\right\}dy=\int_0^1\left\{\frac{-1}{1+y^2}\right\}dy=\frac{-\pi}{4}$$
$$\int_0^1\left\{\int_0^1\frac{y^2-x^2}{(x^2+y^2)^2}\,dy\right\}dx=\int_0^1\left\{\frac{+1}{1+x^2}\right\}dx=\frac{+\pi}{4}.$$

As expected in light of Fubini's Theorem, the integrand isn't nonnegative nor is it in L_1:

$$\begin{aligned}\iint_{0,0}^{1,1}\left|\frac{y^2-x^2}{(x^2+y^2)^2}\right|dx\,dy&\ge\int_0^{\pi/2}\int_0^1\frac{r^2|\sin^2\theta-\cos^2\theta|}{r^4}\,r\,dr\,d\theta\\&=\left(\int_0^{\pi/2}|\cos(2\theta)|\,d\theta\right)\left(\int_0^1 r^{-1}\,dr\right)\\&=(1)\,(\infty)\,.\end{aligned}$$

6.8.2 A Simple but Useful Consequence of Fubini

For any $p>0$ and any random variable X,

$$\begin{aligned}\mathsf{E}|X|^p=\mathsf{E}\left[\int_0^{|X|}p\,x^{p-1}\,dx\right]&=\mathsf{E}\left[\int_0^{\infty}\mathbf{1}_{\{|X|>x\}}p\,x^{p-1}\,dx\right]\\=\int_0^{\infty}\left[\mathsf{E}\mathbf{1}_{\{|X|>x\}}\right]p\,x^{p-1}\,dx&=\int_0^{\infty}p\,x^{p-1}\mathsf{P}\left[|X|>x\right]dx.\end{aligned}$$

It follows that

$$X\in L_p\Leftrightarrow\mathsf{E}|X|^p<\infty\Leftrightarrow\sum_{n=1}^{\infty}n^{p-1}\mathsf{P}[|X|>n]<\infty.$$

To see this, set $Y := |X|$ and fix $p \geq 1$. Then, since $\lceil \frac{n+1}{n} \rceil \leq 2$ for $n \in \mathbb{N}$, for $p \geq 1$:

$$\begin{aligned}
\mathsf{E}Y^p \leq \mathsf{E}\big(\lceil Y \rceil^p\big) &= \int_0^\infty p\, y^{p-1} \mathsf{P}\big[\lceil Y \rceil > y\big]\, dy \\
&= \sum_{n=0}^\infty \int_n^{n+1} p\, y^{p-1} \mathsf{P}\big[\lceil Y \rceil > y\big]\, dy \\
&= \sum_{n=0}^\infty \int_n^{n+1} p\, y^{p-1} \mathsf{P}\big[Y > n\big]\, dy \\
&\leq \sum_{n=0}^\infty p\,(n+1)^{p-1} \mathsf{P}\big[Y > n\big] \\
&\leq p + \sum_{n=1}^\infty p\, 2^{p-1} n^{p-1} \mathsf{P}\big[Y > n\big]
\end{aligned}$$

and hence $Y \in L_p$ if $\sum n^{p-1} \mathsf{P}[Y > n]$ converges. Conversely, if $0 \leq Y \in L_p$ for $p \geq 1$, then:

$$\begin{aligned}
\mathsf{E}Y^p \geq \mathsf{E}\big(\lfloor Y \rfloor^p\big) &= \int_0^\infty p\, y^{p-1} \mathsf{P}\big[\lfloor Y \rfloor > y\big]\, dy \\
&= \sum_{n=0}^\infty \int_n^{n+1} p\, y^{p-1} \mathsf{P}\big[Y > n\big]\, dy \\
&\geq \sum_{n=0}^\infty \int_n^{n+1} p\, n^{p-1} \mathsf{P}\big[Y > n\big]\, dy \\
&= \sum_{n=1}^\infty p\, n^{p-1} \mathsf{P}\big[Y > n\big],
\end{aligned}$$

so the sum converges if $Y \in L_p$. The argument for $0 < p < 1$ is very similar, but differs slightly because now y^{p-1} is decreasing on each interval $(n, n+1]$ instead of increasing.

The case $p = 1$ is easiest and most important: if $S := \sum_{n=0}^\infty \mathsf{P}[|X| > n] < \infty$, then $X \in L_1$ with $\mathsf{E}|X| \leq S \leq \mathsf{E}|X| + 1$. If X takes on only nonnegative integer values then $\mathsf{E}X = S$. For any $\epsilon > 0$, apply this to $Y := X/\epsilon$ to see

$$\epsilon \sum_{n=1}^\infty \mathsf{P}[|X| > n\epsilon] < \mathsf{E}|X| \leq \epsilon \sum_{n=0}^\infty \mathsf{P}[|X| > n\epsilon]$$

6.9 Hoeffding's Inequality

If $\{X_j\}$ are independent and (individually) bounded, so $(\forall\ j \in \mathbb{N})\ (\exists\ \{a_j, b_j\})$ for which $\mathsf{P}[a_j \le X_j \le b_j] = 1$, then $(\forall c > 0),\ S_n := \sum_{j=1}^n X_j$ satisfies

$$\mathsf{P}\big[(S_n - \mathsf{E}S_n) \ge c\big] \le \exp\Big(-2c^2 \Big/ \sum_1^n |b_j - a_j|^2 \Big).$$

If X_j are iid and bounded by $\|X_j\|_\infty \le 1$, e.g., then take $a_j = -1$, $b_j = 1$, and $c = n\epsilon$ to see

$$\mathsf{P}\big[(\bar X_n - \mu) \ge \epsilon\big] \le e^{-n\epsilon^2/2}. \tag{2}$$

Wassily Hoeffding proved this improvement on Chebychev's inequality for L_∞ random variables in 1963 at UNC. It follows from Hoeffding's Lemma:

$$\mathsf{E}[e^{\lambda(X_j - \mu_j)}] \le \exp\big(\lambda^2 (b_j - a_j)^2/8\big),$$

proved in turn from Jensen's ineq and Taylor's theorem (with remainder). The importance of (2) is that the bound decreases *exponentially* in n as $n \to \infty$, while the Chebychev bound $\mathsf{P}\big[|\bar X_n - \mu| \ge \epsilon\big] \le \sigma^2/n\epsilon^2$ decreases only like $1/n$. The price for this better bound is that the $\{X_j\}$ must be bounded in L_∞, not merely in L_2. See also related and earlier **Bernstein**'s inequality (1937), **Chernoff** bounds (1952), and **Azuma**'s inequality (1967).

Here's a proof for the important special case of $X_j = \pm 1$ with probability 1/2 each (and hence $\mu = 0$):

$$\begin{aligned}
\mathsf{P}[\bar X_n \ge \epsilon] &= \mathsf{P}[S_n \ge n\epsilon] \\
&= \mathsf{P}\left[e^{\lambda S_n} \ge e^{n\lambda\epsilon}\right] && \text{for any } \lambda > 0 \\
&\le \mathsf{E}\ e^{\lambda S_n}\ e^{-n\lambda\epsilon} && \text{by Markov's inequality} \\
&= \left\{\tfrac12 e^{\lambda} + \tfrac12 e^{-\lambda}\right\}^n\ e^{-n\lambda\epsilon} && \text{by independence} \\
&\le \left\{e^{\lambda^2/2}\right\}^n\ e^{-n\lambda\epsilon} && \text{see footnote}^3 \\
&= \exp\big(n\lambda^2/2 - n\lambda\epsilon\big)
\end{aligned}$$

The exponent is minimized at $\lambda = \epsilon$, so the tightest bound is:

$$\mathsf{P}[\bar X_n \ge \epsilon] \le \exp\big(n\epsilon^2/2 - n\epsilon\epsilon\big) = e^{-n\epsilon^2/2}.$$

The general case isn't much harder, but proving $\mathsf{E}e^{\lambda(X-\mu)} \le e^{\lambda^2/2}$ is a bit more delicate.

By Borel/Cantelli it follows from Hoeffding's inequality that $(\bar X_n - \mu) > \epsilon$ only finitely-many times for each $\epsilon > 0$, if $\{X_n\} \subset L_\infty$ are iid, leading to our first **Strong Law of Large Numbers**: $\mathsf{P}[\bar X_n \to \mu] = 1$ (why does this follow?).

[3] $\cosh(\lambda) = \left\{\tfrac12 e^{\lambda} + \tfrac12 e^{-\lambda}\right\} = \sum \frac{\lambda^{2k}}{(2k)!} \le \sum \frac{(\lambda^2)^k}{2^k (k!)} = e^{\lambda^2/2}$

Note that Chebychev's inequality only guarantees the algebraic bound $\mathsf{P}[\bar{X}_n \geq \epsilon] \leq 1/n\epsilon^2$, instead of Hoeffding's exponential bound. Since $1/n\epsilon^2$ isn't summable in n, Chebychev's bound isn't strong enough to prove a strong LLN, but Hoeffding's is.

Hoeffding's inequality is now used commonly in computer science and machine learning, applied to indicators of Bernoulli events (or, equivalently, to binomial random variables). It gives the bound

$$\mathsf{P}\big[|\bar{X}_n - p| \leq \epsilon\big] = \mathsf{P}\big[(p-\epsilon)n \leq S_n \leq (p+\epsilon)n\big] \geq 1 - 2e^{-2\epsilon^2 n}$$

for $S_n := \sum_{j\leq n} X_j \sim \mathsf{Bi}(n,p)$ for iid Bernoulli $X_j \overset{\text{iid}}{\sim} \mathsf{Bi}(1,p)$ variables, showing exponential concentration of probability around the mean. This is far stronger than Chebychev's bound of $\mathsf{P}\big[|\bar{X}_n - p| \leq \epsilon\big] \geq 1 - p(1-p)/\epsilon^2 n$, since Hoeffding's bound for $\mathsf{P}\big[|\bar{X}_n - p| > \epsilon\big]$ is *exponentially* small as $n \to \infty$ and hence is summable.