Stat 501 – Probability Theory I
Some Lecture Notes

Ryan Martin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
rgmartin@uic.edu
www.math.uic.edu/~rgmartin

October 20, 2014

Abstract

These notes are based on the two lectures[1] I gave in Stat 501, *Probability Theory I*, as a substitute for Professor Cheng Ouyang. There is a brief discussion of the so-called "basic grouping lemma" concerning independent σ-algebras, followed by a more detailed discussion of the Borel–Cantelli lemma, some applications, and some elaborations on independence.

Independence, and the basic grouping lemma

Let $\{\mathcal{B}_t : t \in T\}$ be a collection of *independent* σ-algebras, i.e., for any $k \geq 1$, for any $t_1, \ldots, t_k$, and any $B_{t_1}, \ldots, B_{t_k}$ in $\mathcal{B}_{t_1}, \ldots, \mathcal{B}_{t_k}$, the events $B_{t_1}, \ldots, B_{t_k}$ are independent. It makes sense that disjoint sub-collections of the σ-algebras are also independent. Here is the formal result.

Lemma (Grouping Lemma). *Let $\{B_t : t \in T\}$ be an independent collection of σ-algebras. Let S be an index set with the property that, for $s \in S$, $T_s \subset T$ and $\{T_s : s \in \mathcal{S}\}$ are pairwise disjoint. Define*

$$\mathcal{B}_{T_s} = \textit{smallest } \sigma\textit{-algebra containing all } \mathcal{B}_t,\ t \in T_s.$$

Then $\{\mathcal{B}_{T_s} : s \in S\}$ is an independent collection of σ-algebras.

Proof. Pretty easy, see pages 101–102 in Resnick. $\square$

Despite the complicated σ-algebra terminology, the Grouping Lemma is quite intuitive. For example, let $X_1, \ldots, X_n$ be a collection of independent random variables. Then the Grouping Lemma says that

- $\sigma(\{X_1, \ldots, X_k\})$ and $\sigma(\{X_{k+1}, \ldots, X_n\})$ are independent σ-algebras;
- $\sum_{i=1}^k X_i$ and $\sum_{i=k+1}^n X_i$ are independent random variables;
- and, more generally, $f(X_1, \ldots, X_k)$ and $g(X_{k+1}, \ldots, X_n)$ are independent random variables for any suitable real-valued functions f and g.

[1]These lectures are based, in part, on Sidney Resnick's *A Probability Path*.

Borel–Cantelli lemma, part I

For a given sequence of events $\{A_n : n \geq 1\}$ in a common probability space, recall the notion of "limsup" of events. That is

$$\limsup_n A_n = \bigcap_{N \geq 1} \bigcup_{n \geq N} A_n =: \{A_n, \text{i.o.}\}.$$

That is, A_n occurs "infinitely often" in the sense that, for ever N, there exists n such that A_n occurs. Then the first Borel–Cantelli lemma concerns the probability of this "infinitely often" event.

Theorem (Borel–Cantelli, part I)**.** *Let A_n be events in the probability space $(\Omega, \mathcal{A}, P)$. If $\sum_n P(A_n) < \infty$, then $P(\limsup_n A_n) = P(A_n, i.o.) = 0$.*

Proof. It is clear that $\limsup_n A_n \subseteq \bigcup_{n \geq N} A_n$ for any N. Furthermore, by monotonicity of P and Boole's inequality $P(\limsup_n A_n) \leq P(\bigcup_{n \geq N} A_n) \leq \sum_{n \geq N} P(A_n)$. Since the summation over all n is finite, given any $\varepsilon > 0$, there exists $N = N_\varepsilon$ such that $\sum_{n \geq N_\varepsilon} P(A_n) < \varepsilon$. Since ε is arbitrary, it follows that $P(\limsup_n A_n) = 0$. □

Application: Strong law of large numbers

The Borel–Cantelli lemma above looks very simple, but has important consequences. In fact, this result is commonly used to prove various almost sure convergence results. Besides the application here, I am aware of many applications of the Borel–Cantelli lemma in proofs of consistency for general Bayesian posterior distributions.

This section gives an example that is a bit ahead of the course schedule. However, since the Borel–Cantelli lemma is so useful, it makes sense to give an interesting application to highlight its importance. Consider a sequence of iid random variables $X_1, X_2, \ldots$, with common mean μ. A result that is justified, but not proved, in an introductory probability course is the law of large numbers, i.e.,

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0, \quad n \to \infty, \quad \forall\, \varepsilon > 0,$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean. That is, for large samples, the sample mean $\bar{X}_n$ will be close to the population mean μ with high probability. This is often referred to as a *weak law of large numbers*. Here, using the Borel–Cantelli lame, we will prove a stronger result, namely, a *strong law of large numbers*,[2] but for only a special case of standard normal random variables.

Let $X_1, X_2, \ldots$ be a sequence of independent $\mathsf{N}(0,1)$ random variables, i.e., their common density function is

$$(2\pi)^{-1/2} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

The crucial step to the application of the Borel–Cantelli lemma and the strong law of large numbers is the following "concentration inequality" for normal random variables.

Lemma. *Let $Z \sim \mathsf{N}(0,1)$. Then, for any $\varepsilon > 0$,*

$$P(|Z| > \varepsilon) \leq \varepsilon^{-1} e^{-\varepsilon^2/2}.$$

[2]The difference between the "weak" and "strong" law of large numbers is the mode of convergence: the former is "in probability," while the latter is "with probability 1." You will learn more about the difference between these two modes of convergence later in Stat 510.

Proof. For $Z \sim \mathsf{N}(0,1)$ and $\varepsilon > 0$, we get

$$P(Z > \varepsilon) = \int_\varepsilon^\infty \frac{1}{(2\pi)^{1/2}} e^{-z^2/2}\, dz \le \frac{1}{(2\pi)^{1/2}} \int_\varepsilon^\infty \frac{z}{\varepsilon} e^{-z^2/2}\, dz.$$

By substitution ($u = z^2/2$), we can simplify the upper bound to get

$$P(Z > \varepsilon) \le \varepsilon^{-1}(2\pi)^{-1/2} e^{-\varepsilon^2/2}.$$

By symmetry of the standard normal,

$$P(|Z| > \varepsilon) = 2P(Z > \varepsilon) \le \frac{2}{(2\pi)^{1/2}} \frac{1}{\varepsilon} e^{-\varepsilon^2/2} \le \varepsilon^{-1} e^{-\varepsilon^2/2}. \qquad \square$$

This can be extended to the case of a sample mean for n independent standard normals, due to the well-known fact that $\bar{X}_n \sim \mathsf{N}(0, n^{-1})$ when $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$.

Corollary. *For all n and all $\varepsilon > 0$, $P(|\bar{X}_n| > \varepsilon) \le n^{-1/2} \varepsilon^{-1} e^{-n\varepsilon^2/2}$.*

Proof. The distribution of $\bar{X}_n$ is the same as $n^{-1/2} Z$, for $Z \sim \mathsf{N}(0,1)$. Then

$$P(|\bar{X}_n| > \varepsilon) = P(|Z| > n^{1/2}\varepsilon).$$

Write $\varepsilon' = n^{1/2}\varepsilon$ and apply the previous lemma. □

The main theorem of this section is a law of large numbers for $\bar{X}_n$, that is, we claim that the sample mean, $\bar{X}_n$, of standard normal random variables converges to the population mean, 0, with probability 1. A key step to proving this claim is representing the event of non-convergence as a lim sup. Given $\varepsilon > 0$, let $A_n = \{|\bar{X}_n| > \varepsilon\}$. Then the event $\{\bar{X}_n \not\to 0\}$ is the same as $\limsup_n A_n$. To understand this, think about convergence using the calculus definition of limit:

> *A sequence of numbers x_n converges to x if, for any $\varepsilon > 0$, there exists $N = N(\varepsilon, x)$ such that $|x_n - x| < \varepsilon$ for all $n > N$.*

In this sense, a sequence x_n does not converge to x if there exists ε such that $|x_n - x| > \varepsilon$ for infinitely many n. This proves that $\{\bar{X}_n \not\to 0\} = \limsup_n A_n$.

Theorem. *For $X_1, X_2, \ldots \overset{iid}{\sim} \mathsf{N}(0,1)$, $P(\bar{X}_n \not\to 0) = 0$.*

Proof. For A_n as defined above, the Corollary gives $P(A_n) \le \varepsilon^{-1} e^{-n\varepsilon^2/2}$. Since $\sum_n P(A_n) < \infty$, it follows from the Borel–Cantelli lemma that $P(\limsup_n A_n) = 0$. From the above equivalence, we get $P(\bar{X}_n \not\to 0) = 0$. □

The above theorem says that $\bar{X}_n$ converges to its mean, 0, with probability 1. This is a stronger result than the weak law of large numbers in Stat 401. The latter only requires a finite variance (which clearly holds for the normal), while the former requires some stronger control on the tails of the normal distribution. It turns out that even the strong law of large numbers holds in broad generality, without needing the Borel–Cantelli lemma and/or concentration inequalities like in the Lemma; in fact, all that's needed is $E|X_1| < \infty$ in the iid case. You will see these details later on in Stat 501/502.

Borel–Cantelli lemma, part II

Note that the first Borel–Cantelli lemma above does not have anything to do with independence. It turns out that there is a second version of the result which says, for independent events A_n, $P(A_n, \text{i.o.})$ is either 0 or 1, depending on whether the series $\sum_n P(A_n)$ converges or diverges.

Theorem (Borel–Cantelli, part II)**.** *Let $\{A_n : n \geq 1\}$ be independent events. Then*

$$P(A_n, i.o.) = \begin{cases} 0 & \text{if } \sum_n P(A_n) < \infty, \\ 1 & \text{if } \sum_n P(A_n) = \infty. \end{cases}$$

Proof. The first part follows from the previous result. For the second part, we show that $[\limsup_n A_n]^c = \liminf_n A_n^c = \bigcup_{N \geq 1} \bigcap_{n \geq N} A_n^c$ has probability 0 if $\sum_n P(A_n) = \infty$. Of course,

$$\liminf_n A_n^c = \bigcup_{N \geq 1} \bigcap_{n \geq N} A_n^c \subseteq \bigcap_{n \geq N} A_n^c, \quad \forall\, N \geq 1.$$

By independence, the set on the right-hand side above has probability

$$P\Bigl(\bigcap_{n \geq N} A_n^c\Bigr) = \prod_{n \geq N} \{1 - P(A_n)\}.$$

The claim is that the right-hand side above equals zero. Pick a finite integer M. Since $1 - x \leq e^{-x}$, we get

$$\prod_{n=N}^{N+M} \{1 - P(A_n)\} \leq \prod_{n=N}^{N+M} e^{-P(A_n)} = e^{-\sum_{n=N}^{N+M} P(A_n)}.$$

Since the series $\sum_n P(A_n)$ diverges, the exponent in the upper bound approaches $-\infty$ as $M \to \infty$, and it follows that $P(\limsup_n A_n) = 1 - P(\liminf_n A_n^c) = 1 - 0 = 1$. □

In applications, at least in statistics, it is rare for the events A_n of interest to be independent. So, as far as I know, the first Borel–Cantelli lemma is the most useful in practice. However, the second Borel–Cantelli lemma provides a first peek at an interesting phenomenon that occurs more generally. That is, for certain "limiting events," the probability can be either 0 or 1, nothing in between. You will see later under the name *Kolmogorov Zero-One Law*, which implies, for example, that a law of large numbers convergence holds with either probability 1 or probability 0, no middle ground. An interesting and somewhat philosophical implication is that there is no uncertainty in the limiting case—this is, effectively, the *fundamental theorem of statistics*.[3]

Resnick considers an independent coin-flipping experiment, and the event of interest is $\limsup_n A_n = \{\text{coin lands heads infinitely often}\}$. If the coin is fixed, so that $P(A_n) \equiv p$, then clearly the series diverges so the coin lands on heads infinitely often with probability 1, which is quite intuitive. On the other hand, in order for heads to not appear infinitely often, the probability $P(A_n)$ for heads on the nth flip must decay quite rapidly.

[3]This is my terminology, not given in a book, etc—but I don't think that makes it wrong!