STA 711: Probability & Measure Theory
Robert L. Wolpert

6 Independence

6.1 Independent Events

A collection of events $\{A_i\} \subset \mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ is called *independent* if

$$\mathsf{P}\left[\cap_{i\in I} A_i\right] = \prod_{i\in I} \mathsf{P}[A_i]$$

for each finite set I of indices. This is a stronger requirement than "pairwise independence," the requirement merely that

$$\mathsf{P}[A_i \cap A_j] = \mathsf{P}[A_i]\mathsf{P}[A_j]$$

for each $i \neq j$. For a simple counter-example, toss two fair coins and let H_n be the event "Heads on the nth toss" for $n = 1, 2$. Then the three events $A_1 := H_1$, $A_2 := H_2$, and $A_3 := H_1 \Delta H_2$ (the event that the coins disagree) each have $\mathsf{P}[A_i] = 1/2$ and each pair has $\mathsf{P}[A_i \cap A_j] = (1/2)^2 = 1/4$ for $i \neq j$, but $\cap A_i = \emptyset$ has probability zero and not $(1/2)^3 = 1/8$.

6.1.1 The Borel-Cantelli Lemmas and Borel's Zero-One Law

Our proof below of the Strong Law of Large Numbers for iid bounded random variables relies on the almost-trivial but very useful:

Lemma 1 (Borel-Cantelli) *Let $\{A_n\}$ be events on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$ that satisfy*

$$\sum_{n=1}^{\infty} \mathsf{P}[A_n] < \infty.$$

Then the event that infinitely-many of the $\{A_n\}$ occur (the $\limsup$*) has probability zero.*

Proof.

$$\mathsf{P}\Big[\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\Big] \le \mathsf{P}\Big[\bigcup_{m=n}^{\infty} A_m\Big] \le \sum_{m=n}^{\infty} \mathsf{P}[A_m] \to 0 \quad \text{as } n \to \infty.$$

□

This result does *not* require independence of the $\{A_n\}$, but its partial converse does:

Lemma 2 (Second Borel-Cantelli) *Let $\{A_n\}$ be **independent** events on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ that satisfy*

$$\sum_{n=1}^{\infty} \mathsf{P}[A_n] = \infty.$$

Then the event that infinitely-many of the $\{A_n\}$ occur (the $\limsup$*) has probability one.*

Proof. First recall that $1 + x \le e^x$ for all real $x \in \mathbb{R}$, positive or not (draw a graph). For each pair of integers $1 \le n \le N < \infty$,

$$\begin{aligned}
\mathsf{P}\Big[\bigcap_{m=n}^{N} A_m^c\Big] &= \prod_{m=n}^{N} \big(1 - \mathsf{P}[A_m]\big) \\
&\le \prod_{m=n}^{N} e^{-\mathsf{P}[A_m]} = \exp\left(-\sum_{m=n}^{N} \mathsf{P}[A_m]\right) \\
&\to \exp\left(-\sum_{m=n}^{\infty} \mathsf{P}[A_m]\right) = e^{-\infty} = 0
\end{aligned}$$

as $N \to \infty$; thus each $\cap_{m=n}^{\infty} A_m^c$ is a null set, hence so is their union, so

$$\begin{aligned}
\mathsf{P}\Big[\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\Big] &= 1 - \mathsf{P}\Big[\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m^c\Big] \\
&\ge 1 - \sum_{n=1}^{\infty} \mathsf{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 1 - 0 = 1.
\end{aligned}$$

□

Together these two results comprise the

Proposition 1 (Borel's Zero-One Law) *For independent events $\{A_n\}$, the event $A := \limsup A_n$ has probability $\mathsf{P}(A) = 0$ or $\mathsf{P}(A) = 1$, depending on whether the sum $\sum \mathsf{P}(A_n)$ is finite or not.*

6.1.2 B/C Illustration

Here's a little toy example to illustrate the Borel-Cantelli lemmas. Begin with a leather bag containing one gold coin, and $n = 1$.

(a) At nth turn, first add one additional *silver* coin to the bag, then draw one coin at random. Let A_n be the event

$$A_n = \{\text{Draw gold coin on } n\text{th draw}\}.$$

Whichever coin you draw, replace it; increment n; and repeat.

(b) As above— but at nth turn, add n silver coins.

Let γ be the probability that you *ever* draw the gold coin. In each case, is $\gamma = 0$? $\gamma = 1$? or $0 < \gamma < 1$? In latter case, give exact asymptotic expression for γ and numerical estimate to four decimals. Why doesn't $0 < \gamma < 1$ violate Borel's zero-one law (Prop. 1 below)? Can you find γ exactly, perhaps with the help of Mathematica or Maple?

6.2 Independent Classes of Events

Classes $\{\mathcal{C}_i\}$ of events (*e.g.*, π-systems or σ-algebras) are called *independent* if

$$\mathsf{P}\left[\cap_{i\in I} A_i\right] = \prod_{i\in I} \mathsf{P}[A_i]$$

for each finite I whenever each $A_i \in \mathcal{C}_i$. An important tool for simplifying proofs of independence is

Theorem 3 (Basic Criterion) *If classes $\{\mathcal{C}_i\}$ of events are independent and if each $\mathcal{C}_i$ is a π-system, then $\{\sigma(\mathcal{C}_i)\}$ are independent too.*

Proof. Let I be a finite index set with at least $|I| \geq 2$ elements and $\{\mathcal{C}_i\}_{i\in I}$ an independent collection of π-systems. Fix $i \in I$, set $J := I \setminus \{i\}$, and fix $A_j \in \mathcal{C}_j$ for each $j \in J$. Set:

Then
$$\mathcal{L} := \left\{B \in \mathcal{F} : \quad \mathsf{P}\Big[B \cap \bigcap_{j\in J} A_j\Big] = \mathsf{P}[B] \cdot \prod_{j\in J} \mathsf{P}[A_j]\right\}.$$

- $\mathcal{C}_i \subset \mathcal{L}$, by the hypothesis that $\{\mathcal{C}_i\}$ are independent;
- $\Omega \in \mathcal{L}$, by the independence of $\{\mathcal{C}_j\}_{j\in J}$;
- $B \in \mathcal{L} \Rightarrow B^c \in \mathcal{L}$, by a quick computation; and
- $B_n \in \mathcal{L}$ and $\{B_n\}$ disjoint $\Rightarrow \cup B_n \in \mathcal{L}$, another quick computation.

Thus $\mathcal{L}$ is a λ-system containing $\mathcal{C}_i$, and so by Dynkin's π-λ theorem it contains $\sigma(\mathcal{C}_i)$. Thus $\sigma(\mathcal{C}_i)$ and $\{A_j\}_{j\in J}$ are independent for each $\{A_j \in \mathcal{C}_j\}$, so $\{\sigma(\mathcal{C}_i), \{\mathcal{C}_j\}_{j\in J}\}$ are independent π-systems. Repeating the same argument $|I| - 1$ times (or, more elegantly, mathematical induction on the cardinality $|I|$) completes the proof. ∎

6.3 Independent Random Variables

A collection of random variables $\{X_i\}$ on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$ are called *independent* if

$$\mathsf{P}\big(\cap_{i\in I}\,[X_i \in B_i]\big) = \prod_{i\in I} \mathsf{P}[X_i \in B_i]$$

for each finite set I of indices and each collection of Borel sets $\{B_i \in \mathcal{B}(\mathbb{R})\}$. This is just the same as the requirement that the σ-algebras $\mathcal{F}_i := \sigma(X_i) = X_i^{-1}(\mathcal{B})$ be independent; by the Basic Criterion it is enough to check that the joint CDFs factor, *i.e.*, that

$$\mathsf{P}\big(\cap_{i\in I}\,[X_i \le x_i]\big) = \prod_{i\in I} F_i(x_i) \tag{1}$$

for each finite index set I and each $x \in \mathbb{R}^I$, or just for a dense set of such x (Why?).

For finitely-many jointly continuous random variables this is equivalent to requiring that the joint density function factor as the product of marginal density functions (proof: differentiate (1) w.r.t. each x_i), while for finitely-many discrete random variables it's equivalent to the usual factorization criterion for the joint pmf. The present definition goes beyond those two cases, however— for example, it includes the case of a discrete random variable $X \sim \mathsf{Bi}(7, 0.3)$, absolutely continuous $Y \sim \mathsf{Ex}(2.0)$, mixed $Z = (\zeta \wedge 0)$ for $\zeta \sim \mathsf{No}(0,1)$, and discrete continuous C with the Cantor distribution. It also applies to infinite (even uncountable) collections of random variables, where no joint pdf or pmf can exist.

Since $\sigma\big(g(X)\big) \subseteq \sigma(X)$ for any random variable X and Borel function $g(\cdot)$, if $\{X_i\}$ are independent and if $g_i(\cdot)$ are arbitrary Borel functions, it follows that $\{g_i(X_i)\}$ are independent too— and, in particular, that if $X \perp\!\!\!\perp Y$ then $X \perp\!\!\!\perp g(Y)$ for all Borel functions $g(\cdot)$. If X and Y are independent, then so are X^2 and $(Y \vee 0)$, for example, with no need to compute joint pdfs or pmfs or the like.

6.4 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 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 4 (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. ∎

6.5 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.

6.6 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 $X \ge 0$ or $\mathsf{E}|X| < \infty$) to guarantee that these three integrals are meaningful and equal:

$$\int_{\Omega_2}\left\{\int_{\Omega_1} X_{\omega_2} d\mathsf{P}_1\right\} d\mathsf{P}_2 \stackrel{?}{=} \iint_{\Omega} X\, d\mathsf{P} \stackrel{?}{=} \int_{\Omega_1}\left\{\int_{\Omega_2} X_{\omega_1} d\mathsf{P}_2\right\} d\mathsf{P}_1 \tag{2}$$

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 (2) 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 $X \ge 0$ 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 5 (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*

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

if either

- *$f(x,y) \ge 0$ 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 (take $\Omega_1 = \mathbb{N}$ with counting measure $\mathsf{P}_1(B_1) := \#\{B_1\}$, and let $K(n, B_2) = \mathsf{P}(X_n \in B_2)$ be the distribution of X_n on $\Omega_2 = \mathbb{R}$), for any sequence of random variables we may exchange summation and expectation and conclude

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

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

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

6.6.1 A Simple but Useful Consequence of Fubini

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

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

It follows that

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

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 $\epsilon > 0$,

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

6.7 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

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

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 is that the bound decreases *exponentially* in n as $n \to \infty$, while the Chebychev bound only decreases like a power of n. The price is that $\{X_j\}$ 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

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

The general case isn't much harder, but proving that $\mathsf{E}e^{\lambda X} \le e^{\lambda^2/2}$ 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.