An Introduction to Laws of Large Numbers

John
CVGMI Group
1 Introduction

Introduction
To Laws of Large Numbers
Weak Law of Large Numbers
Strong Law
Strongest Law

Examples
Information Theory
Statistical Learning

Appendix
Random Variables
Working with R.V.’s
Independence
Limits of Random Variables
Modes of Convergence
Chebyshev
1. INTRODUCTION

2. Introduction To Laws of Large Numbers
   - Weak Law of Large Numbers
   - Strong Law
   - Strongest Law

Examples
- Information Theory
- Statistical Learning

Appendix
- Random Variables
- Working with R.V.’s
- Independence
- Limits of Random Variables
- Modes of Convergence
- Chebyshev
Contents

1 INTRODUCTION

2 Introduction To Laws of Large Numbers
   - Weak Law of Large Numbers
   - Strong Law
   - Strongest Law

3 Examples
   - Information Theory
   - Statistical Learning
We’re working with random variables. What could we observe?
Random Variables \( \{X_n\}_{n=1}^{\infty} \)
We’re working with random variables. What could we observe? Random Variables $\{X_n\}_{n=1}^{\infty}$ ... More specifically

- Bernoulli sequence, looking at probability of average of the sum of $N$ random events. $P(x_1 < \sum_{i=1}^{N} X_i < x_2)$
We’re working with random variables. What could we observe? Random Variables $\{X_n\}_{n=1}^{\infty}$ ... More specifically

- Bernoulli sequence, looking at probability of average of the sum of $N$ random events. $P(x_1 < \sum_{i=1}^{N} X_i < x_2)$
- Coin Flipping anybody?
Intuition

We’re working with random variables. What could we observe? Random Variables \( \{X_n\}_{n=1}^{\infty} \) ... More specifically

- Bernoulli sequence, looking at probability of average of the sum of \( N \) random events. \( P(x_1 < \sum_{i=1}^{N} X_i < x_2) \)
- Coin Flipping anybody?

as \( N \) increases we see that the probability of observing an equal amount of 0 or 1 is \( \sim \) equal
Intuition

We’re working with random variables. What could we observe? Random Variables \( \{X_n\}_{n=1}^{\infty} \) ... More specifically

- Bernoulli sequence, looking at probability of average of the sum of N random events. \( P(x_1 < \sum_{i=1}^{N} X_i < x_2) \)
- Coin Flipping anybody?
  as N increases we see that the probability of observing an equal amount of 0 or 1 is \( \sim \) equal
- This is intuitive: As the number of samples increases the average observation should tend toward the theoretical mean
Let’s try to work out this most basic *fundamental theorem* of random variables arising from repeatedly observing a random event. How do we build such a theorem?
Let’s try to work out this most basic \textit{fundamental theorem} of random variables arising from repeatedly observing a random event. How do we build such a theorem?

- In blackbox scenario we want to be able to use independence.
Let’s try to work out this most basic *fundamental theorem* of random variables arising from repeatedly observing a random event. How do we build such a theorem?

- In blackbox scenario we want to be able to use independence.
- We want to be able to use variance and expectation: so let’s make $X_n \in L^2$ for all $n$, and associate with each $X_n$ it’s mean, we’ll denote by $\mu_n$, and variance, by $\sigma_n$. 
Let’s try to work out this most basic *fundamental theorem* of random variables arising from repeatedly observing a random event. How do we build such a theorem?

- In blackbox scenario we want to be able to use independence.
- We want to be able to use variance and expectation: so let’s make $X_n \in L^2$ for all $n$, and associate with each $X_n$ its mean, we’ll denote by $\mu_n$, and variance, by $\sigma_n$.
- We will create new random variables from the sequence and work with these; aim for results *in terms of* them.
Let’s try to work out this most basic fundamental theorem of random variables arising from repeatedly observing a random event. How do we build such a theorem?

- In blackbox scenario we want to be able to use independence.

- We want to be able to use variance and expectation: so let’s make $X_n \in L^2$ for all $n$, and associate with each $X_n$ its mean, we’ll denote by $\mu_n$, and variance, by $\sigma_n$.

- We will create new random variables from the sequence and work with these; aim for results in terms of them.

- This seems like a start, let’s try to prove a theorem...
Weak Law

Theorem

Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma_n \). Then \( \sum_{i=1}^{n} X_i - \mu_i \to 0 \).

So a sequence of functions on \( \Omega \), the sample space, are going towards a function \( = \) zero...
Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma_n \). Then \( \sum_{i=1}^{n} X_i - \mu_i \to 0 \).

So a sequence of functions on \( \Omega \), the sample space, are going towards a function \( = \) zero...

- Can we prove this?
Weak Law

Theorem

Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma_n \). Then \( \sum_{i=1}^{n} X_i - \mu_i \to 0 \).

So a sequence of functions on \( \Omega \), the sample space, are going towards a function = zero...

- Can we prove this?
- We haven’t used \( \sigma_n \): we’ll clearly need these since otherwise the \( X_i \) are wildly unpredictable.
**Weak Law**

---

**Theorem**

Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma_n \). Then \( \sum_{i=1}^{n} X_i - \mu_i \to 0 \).

So a sequence of functions on \( \Omega \), the sample space, are going towards a function \( = \) zero...

- Can we prove this?
- We haven’t used \( (\sigma_n) \): we’ll clearly need these since otherwise the \( X_i \) are wildly unpredictable.

- IDEA: Constrain \( \lim \sum_{i=1}^{n} \sigma_i^2 = 0 \)
Counterexample
Weak Law: Pitfalls

- We haven’t specified a *mode of convergence*: this is a technical point, but if you recall uniform, pointwise, a.e., normed convergence from analysis then the goal of the proof can change radically.

  - IDEA: We want our rule to hold with high probability, that is it should hold for essentially all values of $\omega \in \Omega$. Convergence in the normed space is pretty ambitious at this point.
Weak Law: Pitfalls

- We haven’t specified a mode of convergence: this is a technical point, but if you recall uniform, pointwise, a.e., normed convergence from analysis then the goal of the proof can change radically.
  - IDEA: We want our rule to hold with high probability, that is it should hold for essentially all values of \( \omega \in \Omega \).
    Convergence in the normed space is pretty ambitious at this point.

- We haven’t specified a rate of convergence.
  - The rate of convergence of the \( \sigma_n \) should imply something about the rate of convergence of the \( X_n - \mu_n \).
  - The \( X_n - \mu_n \) should converge slower than the \( \sigma_n \).
Weak Law: Pitfalls

- We haven’t specified a *mode of convergence*: this is a technical point, but if you recall uniform, pointwise, a.e., normed convergence from analysis then the goal of the proof can change radically.
  - IDEA: We want our rule to hold with high probability, that is it should hold for essentially all values of $\omega \in \Omega$. Convergence in the normed space is pretty ambitious at this point.
- We haven’t specified a *rate* of convergence.
  - The rate of convergence of the $\sigma_n$ should imply something about the rate of convergence of the $X_n - \mu_n$.
  - The $X_n - \mu_n$ should converge slower than the $\sigma_n$.
  - IDEA: Weaken constraint to $\lim n^{-2} \sum_{i=1}^{n} \sigma_i^2 = 0$
Weak Law: Pitfalls

- We haven’t specified a *mode of convergence*: this is a technical point, but if you recall uniform, pointwise, a.e., normed convergence from analysis then the goal of the proof can change radically.
  - IDEA: We want our rule to hold with high probability, that is it should hold for essentially all values of $\omega \in \Omega$. Convergence in the normed space is pretty ambitious at this point.

- We haven’t specified a *rate* of convergence.
  - The rate of convergence of the $\sigma_n$ should imply something about the rate of convergence of the $X_n - \mu_n$.
  - The $X_n - \mu_n$ should converge slower than the $\sigma_n$.
  - IDEA: Weaken constraint to $\lim n^{-2} \sum_{i=1}^{n} \sigma_i^2 = 0$
  - The $X_n - \mu_n$ should converge *on average*
Theorem

Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma^2_n \). Then \( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu_i \to 0 \) in measure if \( \lim \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2_i = 0 \).

Let \( Y_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} (X_i(\omega) - \mu_i) \). Then \( E(Y_n) = 0 \) and \( \sigma^2(Y_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2_i \) by independence. Now we can use the limit of \( \sigma^2(Y_n) \) from the hypothesis, and need to prove

\[
P(\{\omega : |Y_n(\omega)| > \epsilon\}) \leq \frac{\sigma^2(Y_n)}{\epsilon^2} \to 0 \quad \text{as} \quad n \to* \infty
\]
Aside: $L^p$ bounds

What does

$$P\left(\{\omega : |Y_n(\omega)| > \epsilon\} \right) \leq \frac{\sigma^2(Y_n)}{\epsilon^2} \to 0 \text{ as } n \to^* \infty$$

actually say?

- Part of $Y_n > \epsilon$ is bounded by the integral of $Y_n^2$ divided by epsilon squared... Thinking about this makes it obvious.
- Better bounds can be obtained: Chernoff bounds
Law of Large Numbers 2.1

**Theorem**

Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma_n^2 \). Then \( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu_i \to 0 \) in measure if \( \lim_{n \to \infty} n^{-2} \sum_{i=1}^{n} \sigma_i^2 = 0. \)

\[ \Downarrow \text{weakened; use } P(\max_k | \sum_{i=1}^{k} X_i - \mu_i | \geq \epsilon) \leq \epsilon^{-2} \sum_{i=1}^{n} \sigma_i^2. \]

**Theorem**

Let \( \{X_n\} \) be a sequence of independent \( L^2 \) random variables with means \( \mu_n \) and variances \( \sigma_n^2 \). Then \( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu_i \to 0 \) a.e. if \( \lim n^{-2} \sum_{i=1}^{n} \sigma_i^2 < \infty. \)
Khinchine’s Law

Theorem

Let \( \{X_n\} \) be a sequence of independent \( L^1 \) random variables, identically distributed, with means \( \mu \). Then \( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu_i \to 0 \) a.e. as \( n \to \infty \).
Now we need to use the laws. Start with an example in source coding.

- \( C : \mathcal{X}(\Omega) \rightarrow \Sigma^* \) encoding events
- Interesting to know how complex \( C \) must be to encode \( (Y_i)_{i=1}^N \)
- Entropy: \( H(x) = E[-\lg(p(x))] \); representation of how uncertain a r.v. is
- Problem: \( p(\cdot) \) is unknown to the function and the distribution needs to be learned. WLLN can be used to answer: how uncertain is the distribution?
### Asymptotic Equipartition Property

**Theorem**

*Asymptotic Equipartition Property*  If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p(x)$ then

$$
\frac{1}{n} \lg(p(x_1, \ldots, x_n)) \rightarrow H(x).
$$

$X_i \overset{i.i.d.}{\sim} p(x)$ so $\lg p(x_i)$ are i.i.d. The weak law of large numbers says that

$$
P\left(| - \frac{1}{n} \left(\prod_i p(x_i)\right) - E\left[\prod_i -\lg(p(x_i))\right]\right| > \epsilon \right) \rightarrow 0
$$

$$
P\left(| - \frac{1}{n} \sum_i \lg((p(x_i))) - E[-\lg(p(x))]\right| > \epsilon \right) \rightarrow 0
$$

so the sample entropy approaches the true entropy in probability as the sample size increases.
Statistical Learning

- Want to minimize $R(f) = E(l(x, y, f(x)))$ eg $l = (x, y, f) \rightarrow 1/2|f(x) - y|$
- Stuck minimizing over all $f$, under a distribution we don’t know... hopeless...
- IDEA: Take Law of Large numbers and apply it to this framework, and hopefully $R_{emp}(f) = 1/m \sum_i l(x_i, y_i, f(x_i)) \rightarrow R(f)$
- Use $L^p$ bounds to prove convergence results on testing error.
Mini Appendix: Measure Theory

I won’t go into too much detail in this regard. If we’ve made it this far then great.

- $(X, \mathcal{A})$ is a measurable space if $\mathcal{A} \subset 2^X$ is closed under countable unions and complements. These correspond to ’measurable events’.

- $(X, \mathcal{A})$ with $\mu$ is a measure space if $\mu : \mathcal{A} \to [0, \infty)$ is countably additive over disjoint sets: $\mu(\cup_i A_i) = \sum_i \mu(A_i)$ if $A_i \cap A_j = \emptyset$ if $i \neq j$.

- More great properties fall out of this quickly: measure of the emptyset is zero, measures are monotonic under the containment (partial) ordering, even dominated and monotone convergence (of sets) come out from this.

- Chebyshev’s Inequality - Let $f \in L^p(R)$ and denote $E_\alpha = \{x : |f| > \alpha\}$ and now $\|f\|_p^p \geq \int_{E_\alpha} |f|^p \geq \alpha^p \mu(E_\alpha)$ by monotonicity and positivity of measures.
L^2 R.V.’s: Definition

We say that a function $f : X \to R$ belongs to the function space $L^p(X)$ if $||f||_p = (\int |f(x)|^p dx)^{1/p} < \infty$ and say $f$ has finite $L^p$ norm.

**Definition**

The fundamental object being considered in statistics is the random variable. Given a measure space $(\Omega, B, \mu) X : \Omega \to \mathbb{R}$ is an $L^2$ random variable if

$$\{X < r\} \in B \quad \forall r \in \mathbb{R}$$

$\mu(\Omega) = 1$ and

$$\left(\int X(\omega)^2 d\mu(\omega)\right)^{1/2} < \infty.$$

We may also say that a random variable with finite 2nd moment is a Borel-measurable real valued function in $L^2(X)$. 
L² R.V.’s: Questions

- Why $\mathcal{B}$?
  - Physical paradoxes, eg. Banach-Tarski
  - Want to talk about physical events $\sim$ measurable sets
  - $\mathcal{B}$ is as descriptive as we’d like most of the time

- What does $\mu$ do here?
  - $\mu$ weights the events $\sim$ measurable sets
  - Puts values to $\{\omega \in \Omega : X(\omega) < r\}$; $\Omega$ is the sample space
  - $X$ maps random events (patterns occurring in data) to the reals which have a good standard notion of measure. So $X$ induces the distribution $P(B) = \mu(X^{-1}(B))$ which gives

$$P(\{X(\omega) \in B\}) = \mu(X^{-1}(B))$$

as expected. This is usually written without the sample variable.

- For more sophisticated questions/answers see MAA 6616 or STA 7466, but discussion is encouraged
Now we can start building

**Theorem**

If \( f : \mathbb{R} \rightarrow \mathbb{R} \) is measurable, then

\[
\int f(x) \, dP = \int f(X(\omega)) \, d\mu
\]

ie. the random variable pushes forward a measure onto \( \mathbb{R} \), and the integrals of measurable functions of random variables are therefore ultimately determined by real integration.

This may be proved using Dirac measures and measure properties. Following these lines we may develop the basics of the theory of probability from measure theory. (ok, Homework)
Define the first moment as the linear functional

\[ E(X) = \int t \, dP. \]

Then the \( p^{th} \) central moment is a functional given by

\[ m_p(X) = \int (t - E(X))^p \, dP. \]

Note that when \( X \in L^2 \) these are well defined by the theorem above. Pullbacks are omitted. Since we’re talking about \( L^2 \) let’s define the second central moment as \( \sigma^2 \) for convenience. If we need higher moments, sadly, mathematics says no.
Now that we know what random variables are, let’s try to define how a collection of them interact in the most basic way possible.

**Definition**

Let $\mathcal{X} = \{X_\alpha\}$ be a collection of random variables. Then $\mathcal{X}$ is independent if

$$P(\bigcap_\alpha X_\alpha^{-1}(B_\alpha)) = P((X_1, \ldots, X_n)^{-1}(B_1 \times \ldots \times B_n))$$

$$= (\mu(X_1^{-1}) \times \ldots \times \mu(X_n^{-1}))(B_1 \times \ldots \times B_n).$$

(Head Explodes) Really this just means that the $X_\alpha$ are independent if their joint distributions are given by the product measure over the induced measures.
The significance of the belabored definition of independence is that when a joint distribution, just a distribution over several random variables, contains zero information about how the variables are related.

- If we have an infinite collection of random variables, independent of each other - in spite of independence - we would like to be able to infer information about the lower order moments from the higher order moments and vice versa.

- If we have ideal conditions on the the random variables then we should be able to deduce information about moments from a very large sequence.

- The type of convergence we get should be sensitive to the hypotheses.
Modes of Convergence

We clearly need to specify how the means are converging, right?
We clearly need to specify how the means are converging, right? Why?
Modes of Convergence

We clearly need to specify *how* the means are converging, right? Why?
So how could \( \sum_{i=1}^{N} \frac{(X_i - \mu_i)}{N} \to 0? \)
We clearly need to specify how the means are converging, right? Why?
So how could $\sum_{i=1}^{N}(X_i - \mu_i)/N \rightarrow 0$?

- In measure: $\lim_{N} P(\{|\sum_{i=1}^{N}(X_i - \mu_i)/N| \geq \epsilon\}) \rightarrow 0$
- In norm: $\lim_{N}(\int (\sum_{i=1}^{N}(X_i - \mu_i)/N)^p d\mu(\omega))^{1/p} \rightarrow 0$
- A.E: $P(\{\lim_{N} |\sum_{i=1}^{N}(X_i - \mu_i)/N| > 0\}) \rightarrow 0$
Chebyshev