1. Interesting Bernoulli Convergence

Consider an independent sequence of random variables where $X_n \sim B\left(\frac{1}{n}\right)$.

(a) Prove that $X_n$ converges to 0 in probability.

(b) Prove that $X_n$ does not converge almost surely to 0.

Solution:

(a) We want to show that for all $\epsilon > 0$, \[ \lim_{n \to \infty} P(|X_n - 0| > \epsilon) = 0. \] We know that $X_n$ can only be 1 or 0 so if $\epsilon > 1$, $P(|X_n - 0| > \epsilon) = 0$. If $0 < \epsilon \leq 1$, we know that $P(|X_n - 0| > \epsilon) = P(X_n = 1) = \frac{1}{n} \cdot \lim_{n \to \infty} \frac{1}{n} = 0$.

(b) Since $X_n$ can only be 1 or 0, for $\lim_{n \to \infty} X_n$ to be 0, there must exist an $N$ such that for all $n \geq N$, $X_n = 0$. We can reword this as either for all $n \geq 1 X_n = 0$ or for all $n \geq 2 X_n = 0$ or for all $n \geq 3 X_n = 0$ or \ldots So

\[
P\left( \lim_{n \to \infty} X_n = 0 \right) = P\left( \bigcup_{N=1}^{\infty} \text{for all } n \geq N X_n = 0 \right)
\leq \sum_{N=1}^{\infty} P(\text{for all } n \geq N X_n = 0)
= \sum_{N=1}^{\infty} P\left( \bigcap_{n=N}^{\infty} X_n = 0 \right)
= \sum_{N=1}^{\infty} \prod_{n=N}^{\infty} P(X_n = 0)
= \sum_{N=1}^{\infty} \frac{N-1}{N} \cdot \frac{N}{N+1} \cdot \frac{N+1}{N+2} \cdot \ldots
= \sum_{N=1}^{\infty} 0
= 0
\]

Since this probability is not 1, $X_n$ does not converge almost surely to 0. In fact, since this probability is 0, $X_n$ almost surely does not converge to 0. This is related to Kolmogorov’s 0-1 Law (covered in Stat 205A), which states that for a sequence of independent random variables $X_n$, $X_n$ either converges or does not converge with probability 1.
2. Convergence in Probability

Let \((X_n)_{n=1}^{\infty}\) be a sequence of i.i.d. random variables distributed uniformly in \([-1, 1]\). Show that the following sequences \((Y_n)_{n=1}^{\infty}\) converge in probability to some limit.

(a) \(Y_n = \prod_{i=1}^{n} X_i\).
(b) \(Y_n = \max\{X_1, X_2, \ldots, X_n\}\).
(c) \(Y_n = (X_1^2 + \cdots + X_n^2)/n\).

Solution:

(a) By independence of the random variables,

\[
E[Y_n] = E[X_1] \cdots E[X_n] = 0,
\]

\[\text{var } Y_n = E[Y_n^2] = (\text{var } X_1)^n = \left(\frac{1}{3}\right)^n.\]

Now since \(\text{var } Y_n \to 0\) as \(n \to \infty\), by Chebyshev’s Inequality the sequence converges to its mean, that is, 0, in probability.

(b) Consider \(\epsilon \in [0, 1]\). We see that:

\[
P(|Y_n - 1| \geq \epsilon) = P(\text{max}\{X_1, \ldots, X_n\} \leq 1 - \epsilon) = P(X_1 \leq 1 - \epsilon, \ldots, X_n \leq 1 - \epsilon) = P(X_1 \leq 1 - \epsilon)^n = \left(1 - \frac{\epsilon}{2}\right)^n.
\]

Thus, \(P(|Y_n - 1| \geq \epsilon) \to 0\) as \(n \to \infty\) and we are done.

(c) The expectation is

\[
E[Y_n] = \frac{1}{n} \cdot n \cdot E[X_1^2] = \frac{1}{3}.
\]

Then, we bound the variance.

\[
\text{var } Y_n = \frac{1}{n} \text{ var } X_1^2 \leq \frac{1}{n} \to 0 \quad \text{as } n \to \infty,
\]

since \(X_1^2 \leq 1\). Hence, we see that \(Y_n \to 1/3\) in probability as \(n \to \infty\).

Remark: We now provide an interpretation for the previous result. The sample space for \(Y_n\) is \(\Omega_n = [-1, 1]^n\), which is an \(n\)-dimensional cube. The result we have just proved shows that, for any \(\epsilon > 0\), the set

\[
B_n = \{x \in \mathbb{R}^n : \frac{1}{3}(1 - \epsilon) \leq \frac{x_1^2 + \cdots + x_n^2}{n} \leq \frac{1}{3}(1 + \epsilon)\}
\]

makes up “most” of the volume of \(\Omega_n\), in the sense that

\[
\frac{\text{volume}(B_n \cap [-1, 1]^n)}{2^n} \to 1 \quad \text{as } n \to \infty.
\]

Since \(B_n\) is close to the boundary of a ball of radius \(\sqrt{n/3}\), the result can be stated facetiously as “nearly all of the volume of a high-dimensional cube is contained in the boundary of a ball”. Although this may seem like a meaningless comment, in fact various phenomena such as these contribute to the so-called “curse of dimensionality” in machine learning, which concerns the sparsity of data in high-dimensional statistics.
3. More Almost Sure Convergence

(a) Suppose that, with probability 1, the sequence \((X_n)_{n \in \mathbb{N}}\) oscillates between two values \(a \neq b\) infinitely often. Is this enough to prove that \((X_n)_{n \in \mathbb{N}}\) does not converge almost surely? Justify your answer.

(b) Suppose that \(Y\) is uniform on \([-1, 1]\), and \(X_n\) has distribution

\[ P(X_n = (y + n^{-1})^{-1} \mid Y = y) = 1. \]

Does \((X_n)_{n=1}^\infty\) converge a.s.?

(c) Define random variables \((X_n)_{n \in \mathbb{N}}\) in the following way: first, set each \(X_n\) to 0. Then, for each \(k \in \mathbb{N}\), pick \(j\) uniformly randomly in \(\{2^k, \ldots, 2^{k+1} - 1\}\) and set \(X_j = 2^k\). Does the sequence \((X_n)_{n \in \mathbb{N}}\) converge a.s.?

(d) Does the sequence \((X_n)_{n \in \mathbb{N}}\) from the previous part converge in probability to some \(X\)? If so, is it true that \(E[X_n] \to E[X]\) as \(n \to \infty\)?

Solution:

(a) Yes. If a sequence oscillates between two values infinitely often, then it does not converge. Here, we have a sequence that oscillates between two values infinitely often (with probability 1), which means that the sequence does not converge (with probability 1). (Perhaps we could name this “almost surely not converging”!)

The above paragraph was very cumbersome to read, which is why we often abbreviate “with probability 1” with a.s. With this abbreviation, here is how the above justification reads: \((X_n)_{n \in \mathbb{N}}\) oscillates between two values infinitely often a.s., so \((X_n)_{n \in \mathbb{N}}\) does not converge a.s.

(b) Yes. Observe that when \(Y = y \neq 0\), \((X_n)_{n \in \mathbb{N}}\) will converge to \(y^{-1}\). When \(Y = 0\), \((X_n)_{n \in \mathbb{N}}\) does not converge; however, \(P(Y = 0) = 0\) since \(Y\) is a continuous random variable. In other words,

\[ P(X_n \text{ does not converge as } n \to \infty) = P(Y = 0) = 0, \]

so \((X_n)_{n \in \mathbb{N}}\) converges a.s.

(c) No. The sequence \((X_n)_{n \in \mathbb{N}}\) oscillates between 0 and successively higher powers of two infinitely often a.s., so it does not converge a.s.

(d) Yes. Fix \(\varepsilon > 0\). For \(n \in \mathbb{Z}_+\), one has

\[ P(|X_n| > \varepsilon) = \frac{1}{2^k}, \]

where \(k = \lfloor \log_2 n \rfloor\). As \(n \to \infty\), the above probability goes to 0, so \(X_n \to 0\) in probability. Intuitively, \((X_n)_{n \in \mathbb{N}}\) has infinitely many oscillations, so it cannot converge a.s. However, the probability of each oscillation shrinks to 0, so \((X_n)_{n \in \mathbb{N}}\) converges in probability.

The expectations do not converge. For all \(n\), one has \(E[X_n] = 1\), so it is not the case that \(E[X_n] \to 0\) as \(n \to \infty\). Hence, convergence in probability is not sufficient to imply that the expectations converge (in fact, almost sure convergence is not sufficient either).
4. Compression of a Random Source

Suppose I’m trying to send a text message to my friend. In general, I know I need \( \log_2(26) \) bits for every letter I want to send because there are 26 letters in the alphabet. However, it turns out if I have some information on the distribution of the letters, I can do better. For example, I might give the letter \( e \) a shorter bit representation because I know it’s the most common. Actually, it turns out the number of bits I need on average is the entropy, and in this problem, we try to show why this is true in general.

Let \( (X_i)_{i=1}^{\infty} \overset{i.i.d.}{\sim} p(\cdot) \), where \( p \) is a discrete PMF on a finite set \( \mathcal{X} \). We know the entropy of a random variable \( X \) is

\[
H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)
\]

Since entropy is really a function of the distribution, we could write the entropy as \( H(p) \).

(a) Show that

\[
-\frac{1}{n} \log_2 p(X_1, \ldots, X_n) \xrightarrow{n \to \infty} H(X_1) \quad \text{almost surely.}
\]

(Here, we are extending the notation \( p(\cdot) \) to denote the joint PMF of \( (X_1, \ldots, X_n) \): \( p(x_1, \ldots, x_n) := p(x_1) \cdots p(x_n) \).)

(b) Fix \( \epsilon > 0 \) and define \( A^{(n)}_\epsilon \) as the set of all sequences \( (x_1, \ldots, x_n) \in \mathcal{X}^n \) such that:

\[
2^{-n(H(X_1)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X_1)-\epsilon)}.
\]

Show that \( \mathbb{P}((X_1, \ldots, X_n) \in A^{(n)}_\epsilon) > 1 - \epsilon \) for all \( n \) sufficiently large. Consequently, \( A^{(n)}_\epsilon \) is called the **typical set** because the observed sequences lie within \( A^{(n)}_\epsilon \) with high probability.

(c) Show that \( (1 - \epsilon)2^{n(H(X_1)-\epsilon)} \leq |A^{(n)}_\epsilon| \leq 2^{n(H(X_1)+\epsilon)} \), for \( n \) sufficiently large. Use the union bound.

Parts (b) and (c) are called the **asymptotic equipartition property (AEP)** because they say that there are \( \approx 2^{nH(X_1)} \) observed sequences which each have probability \( \approx 2^{-nH(X_1)} \). Thus, by discarding the sequences outside of \( A^{(n)}_\epsilon \), we need only keep track of \( 2^{nH(X_1)} \) sequences, which means that a length-\( n \) sequence can be compressed into \( \approx nH(X_1) \) bits, requiring \( H(X_1) \) bits per symbol.

(d) (**optional**) Now show that for any \( \delta > 0 \) and any positive integer \( n \), if \( B_n \subseteq \mathcal{X}^n \) is a set with \( |B_n| \leq 2^{nH(X_1)-\delta} \), then \( \mathbb{P}((X_1, \ldots, X_n) \in B_n) \to 0 \) as \( n \to \infty \).

This says that we cannot compress the observed sequences of length \( n \) into any set smaller than size \( 2^{nH(X_1)} \).

[\text{Hint: Consider the intersection of} \ B_n \text{ and} \ A^{(n)}_\epsilon.\]
(e) \textbf{(optional)} Next we turn towards using the AEP for compression. Recall that in order to encode a set of size \( n \) in binary, it requires \( \lceil \log_2 n \rceil \) bits. Therefore, a naïve encoding requires \( \lceil \log_2 |\mathcal{X}| \rceil \) bits per symbol.

From (b) and (d), if we use \( \log_2 |A_k^{(n)}| \approx nH(X_1) \) bits to encode the sequences in \( A_k^{(n)} \), ignoring all other sequences, then the probability of error with this encoding will tend to 0 as \( n \to \infty \), and thus an asymptotically error-free encoding can be achieved using \( H(X_1) \) bits per symbol. Alternatively, we can create an error-free code by using \( 1 + \lceil \log_2 |A_k^{(n)}| \rceil \) bits to encode the sequences in \( A_k^{(n)} \) and \( 1 + n \lceil \log_2 |\mathcal{X}| \rceil \) bits to encode other sequences, where the first bit is used to indicate whether the sequence belongs in \( A_k^{(n)} \) or not. Let \( L_n \) be the length of the encoding of \( X_1, \ldots, X_n \) using this code; show that \( \lim_{n \to \infty} E[L_n]/n \leq H(X_1) + \epsilon \). In other words, asymptotically, we can compress the sequence so that the number of bits per symbol is arbitrary close to the entropy.

**Solution:**

(a) Since \( (X_i)_{i=1}^\infty \) is an i.i.d. sequence, so is \( (\log_2 p(X_i))_{i=1}^\infty \). Thus:

\[- \frac{1}{n} \log_2 p(X_1, \ldots, X_n) = - \frac{1}{n} \sum_{i=1}^{n} \log_2 p(X_i) \xrightarrow{n \to \infty} a.s. - E[\log_2 p(X_1)] = H(X_1)\]

by the Strong Law of Large Numbers.

(b) As a consequence of (a), \( n^{-1} \log_2 p(X_1, \ldots, X_n) \to H(X_1) \) in probability as \( n \to \infty \), so

\[\mathbb{P}\left( \left| - \frac{1}{n} \log_2 p(X_1, \ldots, X_n) - H(X_1) \right| < \epsilon \right) \to 1 \quad \text{as } n \to \infty.\]

For \( n \) sufficiently large, the LHS is \( > 1 - \epsilon \).

(c) We have:

\[1 = \sum_{x \in \mathcal{X}^n} p(x) \geq \sum_{x \in A_k^{(n)}} p(x) \geq \sum_{x \in A_k^{(n)}} 2^{-n(H(X_1)+\epsilon)} = |A_k^{(n)}| 2^{-n(H(X_1)+\epsilon)} \]

This shows that \( |A_k^{(n)}| \leq 2^{n(H(X_1)+\epsilon)} \). Now, we have, for \( n \) sufficiently large:

\[1 - \epsilon < \mathbb{P}\left( (X_1, \ldots, X_n) \in A_k^{(n)} \right) \leq \sum_{x \in A_k^{(n)}} 2^{-n(H(X_1)-\epsilon)} = 2^{-n(H(X_1)-\epsilon)} |A_k^{(n)}| \]

Thus, \( |A_k^{(n)}| \geq (1 - \epsilon)2^{n(H(X_1)-\epsilon)} \).

(d) Pick \( \epsilon \in (0, \delta) \). We can write

\[\mathbb{P}\left( (X_1, \ldots, X_n) \in B_n \right) \]
\[
\leq \mathbb{P}\left((X_1, \ldots, X_n) \in A^{(n)}_\epsilon \cap B_n\right) + \mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right)
\leq \sum_{x \in A^{(n)}_\epsilon \cap B_n} p(x) + \mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right)
\leq |B_n|2^{-n(H(X_1) - \epsilon)} + \mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right)
\leq 2^{-n(\delta - \epsilon)} + \mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right) \to 0
\]

since \(\delta > \epsilon\) and by (b).

(e) Separating out the sequences in the typical set from the sequences which are not in the typical set,

\[
\frac{\mathbb{E}[L_n]}{n} = \frac{1 + \lceil \log_2 |A^{(n)}_\epsilon| \rceil}{n} \mathbb{P}\left((X_1, \ldots, X_n) \in A^{(n)}_\epsilon\right)
+ \frac{1 + n \lceil \log_2 |X| \rceil}{n} \mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right)
\leq \frac{1 + n[H(X_1) + \epsilon]}{n} + \left(1 + \lceil \log_2 |X| \rceil\right) \mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right).
\]

Since \(\mathbb{P}\left((X_1, \ldots, X_n) \in A^{(n)}_\epsilon\right) \to 1\) and \(\mathbb{P}\left((X_1, \ldots, X_n) \notin A^{(n)}_\epsilon\right) \to 0\), then the second term \(\to 0\). Asymptotically, only the first term matters, and by taking \(n \to \infty\) we get \(\lim_{n \to \infty} \mathbb{E}[L_n]/n \leq H(X_1) + \epsilon\).