AEP as Weak Law of Large Numbers

A theorem from information theory that is a simple consequence of the weak law of large numbers. It states that if a set of values $X_1, X_2, \ldots, X_n$ is drawn independently from a random variable X distributed according to $P(x)$, then the joint probability $P(X_1, \ldots, X_n)$ satisfies

$$-\frac{1}{n}\log_2 P(X_1, X_2, \ldots, X_n) \to H(X),$$

where $H(X)$ is the entropy of the random variable X.

Consider all possibilities of the messages, as you take all possibilities and take their joint probability, it should convergence (in probability) sof joint probability to the entropy of the random variable (similar to convergence in probability of sample average to the mean as one takes many realisations of a random variable).

2.2.2 Law of Large Numbers (Weak)

The standard presentation is as follows. Let $X_1, X_2, \ldots X_n$ be an infinite sequence of independent and identically distributed (Lebesgue integrable) random variables with expected value $\mathbb{E}(X_n) = \mu$ (though one can somewhat relax the i.i.d. assumptions). The sample average $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ converges to the expected value, $\overline{X}_n \to \mu$, for $n \to \infty$.

Finiteness of variance is not necessary (though of course the finite higher moments accelerate the convergence).

The strong law is discussed where needed.

The weak LLN The weak law of large numbers (or Kinchin's law, or sometimes called Bernouilli's law) can be summarized as follows: the probability of a variation in excess of some threshold for the average becomes progressively smaller as the sequence progresses. In estimation theory, an estimator is called consistent if it thus converges in probability to the quantity being estimated.

$$\overline{X}_n \xrightarrow{P} \mu \text{ when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| > \varepsilon \right) = 0.$$

Note that standard proofs are based on Chebyshev's inequality: if X has a finite non-zero variance σ^2. Then for any real number $k > 0$,

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

The use of law of large numbers is to wash out noise also.

of such statistical law, LLN for short: as your sample size grows larger, the anecdote washes out, and the signal progressively takes over the noise. But many –I mean many, if not almost all –psychologists make the reverse mistake: to go from the general to the particular means reestablishing noise!

> *One can only apply particular properties to the general if and only if there is limited noise.*

principle.

'Asymptotic equipartition' principle. For an ensemble of N independent identically distributed (i.i.d.) random variables $X^N \equiv (X_1, X_2, \ldots, X_N)$, with N sufficiently large, the outcome $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is almost certain to belong to a subset of $\mathcal{A}_X^N$ having only $2^{NH(X)}$ members, each having probability 'close to' $2^{-NH(X)}$.

Notice that if $H(X) < H_0(X)$ then $2^{NH(X)}$ is a *tiny* fraction of the number of possible outcomes $|\mathcal{A}_X^N| = |\mathcal{A}_X|^N = 2^{NH_0(X)}$.

The term equipartition is chosen to describe the idea that the members of the typical set have *roughly equal* probability. [This should not be taken too literally, hence my use of quotes around 'asymptotic equipartition'; see page 83.]

A second meaning for equipartition, in thermal physics, is the idea that each degree of freedom of a classical system has equal average energy, $\frac{1}{2}kT$. This second meaning is not intended here.