Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter — Lecture 3
Scribed by Sleiman Itani — T 2/13

In this lecture we will study some convergence results, keeping in our mind that our real objective is to know how well we can approximate a distribution from the data.

1 Probability Convergence Results:

1.1 Axioms of Probability

A *probability space* is defined by a triplet $(\Omega, \mathscr{F}, \mathbf{P})$ where:

1. Ω is the sample space, a collection of all elements.
2. $\mathscr{F}$ is a collection of subsets of Ω that is a $\sigma - field$. It is the collection of events that we are interested in. A $\sigma - field$ $\mathscr{F}$ satisfies the following conditions:
   (a) $\Phi \in \mathscr{F}$
   (b) If $\forall i \in \mathbb{N}$, $\omega_i \in \mathscr{F}$ then $\cup_{i=1}^{\infty} \omega_i \in \mathscr{F}$
   (c) If $\omega \in \mathscr{F}$ then $\omega^c \in \mathscr{F}$
3. $\mathbf{P}$ is a probability measure on $\mathscr{F}$, i.e. $\mathbf{P} : \mathscr{F} \to [0, 1]$ such that:
   (a) $\mathbf{P}(\Phi) = 0$, $\mathbf{P}(\Omega) = 1$
   (b) If $\forall i \in \mathbb{N}$, $\omega_i \in \mathscr{F}$ and $\forall i \neq j$, $\omega_i \cap \omega_j = \Phi$, then:

$$\mathbf{P}(\bigcup_{i=1}^{\infty} \omega_i) = \sum_{i=1}^{\infty} \mathbf{P}(\omega_i)$$

1.2 Random Variables

A random variable is a mapping $x : \Omega \to \mathbf{R}$ (we will sometimes have $x \in [0, 1]$). A sequence of RV's is usually called a random process.

$F(x) = \mathbf{P}\{\omega | X(\omega) \leq x\}$ is the cumulative distribution function. We are interested in sequences of RV's or samples from a certain distribution and asking if they converge to the right place.

1.3 Types of Convergence

Let X_n be a sequence of random variables. The following two types of convergence are of interest to us:

1. $X_n \xrightarrow{p} X$, X_n converges to X in probability if:

$$\forall \epsilon > 0,\ \mathbf{P}\{\omega|\ ||X_n(\omega) - X(\omega)|| > \epsilon\} \to 0\ as\ n \to \infty$$

2. $X_n \xrightarrow{a.s.} X$, X_n converges to X almost surely if:

$$\mathbf{P}\{\omega|\ \lim_{n \to \infty} X_n(\omega) \neq X(\omega)\} = 0$$

Note that both kinds of convergence are very dependent on the measure used and that almost sure convergence implies convergence in probability. To see an example of a sequence that converges in probability and not almost surely, consider the following sequence X_n of independent variables:

$$X_n = \begin{cases} 1 & \text{with probability } n^{-1} \\ 0 & \text{with probability } 1 - n^{-1} \end{cases}$$

It is easy to see that $X_n \xrightarrow{p} 0$ since:

$$\forall \epsilon > 0,\ \mathbf{P}\{\omega |\ ||X_n(\omega)|| > \epsilon\} = n^{-1} \to 0\ as\ n \to \infty$$

But this sequence does not converge almost surely, because $\forall\ 0 < \epsilon < 1$:

$$\mathbf{P}\{\omega |\ X_n(\omega) < \epsilon\ \forall\ n \geq m\} = (1 - m^{-1})(1 - (m+1)^{-1})...$$
$$= \lim_{M \to \infty} \left(\frac{m-1}{m}\right)\left(\frac{m}{m+1}\right)\cdots\left(\frac{M}{M+1}\right) = \lim_{M \to \infty} \frac{m-1}{M+1} = 0\ \forall\ m$$

This means that for all m, the probability that the sequesnce enters the ϵ ball forever at (or before) m is zero, and thus the probability that the sequence converges to 0 is 0.

2 Laws of Large Numbers

In this section we consider a sequence of i.i.d. random variables $X_1, X_2, ...$ such that $X_i : \Omega \to [0,1]$. Here we assume that the sequence of random variables is produced from a sequence of trials from one space that were mapped to $[0,1]$ by the same mapping, i.e. $X_i = X(\omega_i),\ \omega_i \in \Omega$. It is equivalent to assume that each trial is from a different space and mapping $X_i = X_i(\omega_i),\ \omega_i \in \Omega_i$.

Week Law of Large Numbers

$$\frac{1}{l}\sum_{l=1}^{\infty} X_i(\omega) \xrightarrow{p} \mathbf{E}[X]$$

Strong Law of Large Numbers

$$\frac{1}{l}\sum_{l=1}^{\infty} X_i(\omega) \xrightarrow{a.s.} \mathbf{E}[X]$$

We now introduce some very interesting inequalities. Risk minimization uses results that are *similar* to these results.

2.1 Hoeffding inequalities

$\forall \epsilon$, the sequence we have satisfies the following inequalities:

$$\mathbf{P}\{\omega |\ \frac{1}{l}\sum_{l=1}^{\infty} X_i(\omega) - \mathbf{E}[X] > \epsilon\} \leq e^{-2l\epsilon^2}$$

$$\mathbf{P}\{\omega |\ \frac{1}{l}\sum_{l=1}^{\infty} X_i(\omega) - \mathbf{E}[X] < -\epsilon\} \leq e^{-2l\epsilon^2}$$

$$\mathbf{P}\{\omega |\ |\frac{1}{l}\sum_{l=1}^{\infty} X_i(\omega) - \mathbf{E}[X]| > \epsilon\} \leq 2e^{-2l\epsilon^2}$$

Recalling some definitions from the previous lecture:

$$R^l_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(x, y, \alpha)$$

and

$$R(\alpha) = \mathbf{E}[L(x, y, \alpha)]$$

We now know, by the SLLN that $\forall\alpha,\ R^l_{emp}(\alpha) \xrightarrow{a.s.} R(\alpha)$. This convergence is point-wise in α; we are interested in a convergence that is uniform in α:

$$\sup_\alpha |R^l_{emp}(\alpha) - R(\alpha)| \xrightarrow{p} 0$$

We have to study the conditions under which the uniform convergence holds.

3 Non-parametric Density Estimation

Let $X^l : X_1, X_2, ..., X_l$, we want to estimate $F(x)$ and $P(x) = \frac{dF(x)}{dx}$. The most intuitive approach is the following: $\forall x$, our estimate for $F(x)$ is number of $x_i's$ that are less than x, normalized by l. More formally, we define

$$\theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

And

$$F_l(x) = \frac{1}{l}\sum_{i=1}^{l} \theta(x - X_i)$$

With these definitions, we can state the Glivenko-Cantelli theorem:

$$\sup_x |F_l(x) - F(x)| \xrightarrow{a.s} 0$$

We even know that the error can be bounded by a decaying exponential. We usually need to estimate not only $F(x)$, but the probabilities of certain sets. Given $A \subseteq [0, 1]$, we need to estimate $P(A) = \int_A dF(x)$. Our estimate is:

$$\nu(X^l, A) = \int_A dF_l(x) = \frac{\#X_i \in A}{l}$$

We know by the SLLN that $\forall A,\ \nu(X^l, A) \xrightarrow{a.s.} P(A)$. Still, this convergence is not uniform in A. A counter example is when the distribution is uniform, and therefore $F(x) = x$, $\forall l,\ A = x_i, i = 1, 2, ..., l$ has $P(A) = 0$ and $\nu(x^l, A) = 1$. Therefore:

$$\sup_{A\in\mathscr{A}} |\nu(x^l, A) - P(A)| = 1$$

Now we need to find collections of sets on which the convergence is uniform. A nontrivial example is:

$$A^x = (0, x)|x \leq 1$$

This is a direct result of the Glivenko-Cantelli theorem.

If the distribution is continuous, ($P(x) = \frac{dF(x)}{dx}$ is well defined everywhere) then $||P_l(x) - P(x)||_1 \to 0\ as\ l \to \infty$.

In this case, we can approximate the risk by the empirical risk:

$$\sup_\alpha |\int L(x, y, \alpha)P_l(x)dx - \int L(x, y, \alpha)P(x)dx| \leq \max_{x,y,\alpha} |L(x, y, \alpha)|\ ||P_l(x) - P(x)||_1$$

Therefore if $||P_l(x) - P(x)||_1 \to 0$ then the risk error $\to 0$.