Stats 300b: Theory of Statistics **Winter 2017**

Lecture 1 – January 10

Lecturer: John Duchi *Scribe: Xiaowei Wang*

Warning: *these notes may contain factual errors*

Reading: VDV Chapter 2.1, 2.2

Outline of lecture 1:

- Administrative basic stuff: As on the syllabus
- Overview of the course
- Basic theory of convergence of random variables
- Probability, Asymptotic Statistics and Distributions

Overview of the course: (In this course, we will be majorly dealing with big data sets, $N \to \infty$)

1. Convergence of random variables, random vectors, estimators and functions.

2. Understanding various notions of optimality and quality of estimators and tests. We will not be talking about admissibility as it is too difficult. What we will try to do in this course is to show that certain estimators are good under specific metrics or to prove that certain estimators are unimprovable.

Part I of the course: Finite dimensional problems and statistic models

Example 1: One example problem is that we have $X_i \overset{\text{iid}}{\sim} P_\theta, X_i \in \mathbb{R}^d$, where d is fixed. We want to understand the estimators of parameter $\theta \in \mathbb{R}^d$ of distribution P_θ. ♣

Part II of the course: Optimality and comparisons of estimators

In this part, we will try to understand when an estimator $\hat{\theta}$ of θ is good or optimal. Also, we will look into how to distinguish P_θ from $P_{\theta+\Delta}$ when Δ is small.

Part III of the course: Infinite dimensional or uniform laws of convergence for random variables

1. Concentration inequalities

2. For functions $F : X \times \theta \to \mathbb{R}$, we will look into how

$$\frac{1}{n}\sum_{i=1}^{n} F(x_i, \theta) \to \mathbb{E}[F(x, \theta)]$$

uniformly in θ.

Backgrounds needed: 1. Stat 300a (not strictly necessary).

2. Probability at stat 310a level.

e.g. Convergence of distribution, Helly Selection Theorem etc.

3. Analysis at Math 171 level.

e.g. Compactness, metric spaces etc.

Basic theory of convergence of random variables:

In this part we will go thourgh basic definitions, Continuous Mapping Theorem and Portmanteau Lemma.

For now, assume $X_i \in \mathbb{R}^d, d < \infty$. We first give the definition of various convergence of random variables.

Definition 0.1. *(Convergence in probability) We call* $X_n \xrightarrow{p} X$ *(sequence of random variables converges to X) if*

$$\lim_{n\to\infty} \mathbb{P}(||X_n - X|| \geq \epsilon) = 0, \ \forall \epsilon > 0$$

In a general metric space, with metrix ρ, the above definition becomes

$$\lim_{n\to\infty} \mathbb{P}(\rho(X_n, X) \geq \epsilon) = 0, \ \forall \epsilon > 0$$

Definition 0.2. *(Weak convergence or convergence in distribution)*

We say

$$X_n \xrightarrow{d} X$$

if for $\forall x \in \mathbb{R}^d$,

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

at all $X \in \mathbb{R}^d$ *such that* $x \to \mathbb{P}(X \leq x)$ *is continuous.*

Note: In the above definition $\mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x_1] \times \cdots \times (-\infty, x_d])$

We also have an alternative definition for convergence in distribution.

Definition 0.3.

$$X_n \xrightarrow{d} X$$

if for all bounded continuous function f,

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$$

Below is a definition of L^p convergence.

Definition 0.4. *(Convergence in the* p^{th} *mean)*

We say that

$$X_n \xrightarrow{L^p} X$$

if

$$\lim_{n\to\infty} \mathbb{E}[||X_n - X||^p] = 0$$

Finally, we give the definition of almost surely convergence for random variables.

Definition 0.5. *(X_n converges almost surely to X)*
We say that

$$X_n \overset{a.s.}{\to} X$$

if

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

i.e.

$$\mathbb{P}(\lim_{n\to\infty} ||X_n - X|| \geq \epsilon) = 0,\ \forall \epsilon > 0$$

Standard implications:

For the various types of convergence above, we have the following relationships.

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{p}{\to} X \Rightarrow X_n \overset{d}{\to} X$$
$$X_n \overset{L^p}{\to} X \Rightarrow X_n \overset{p}{\to} X$$

All the reversed directions may not be true.

Examples of almost surely convergence and convergence in probability can be found in the strong law of large numbers and central limits theorem, as stated below.

Example 2: If $X_i \overset{\text{iid}}{\sim} P$, $cov(X_i) = \Sigma = \mathbb{E}[(X_i - \mu)(X_i - \mu)^T]$, $\mu = \mathbb{E}[X_i]$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \overset{d}{\to} \mathsf{N}(0, \Sigma)$$

♣

Basic Convergence Theorems: (Chapter VDV for all proofs)

Theorem 1. *(Continuous Mapping Theorem) Let g be continuous on a set B such that $\mathbb{P}(X \in B) = 1$ then*

$$X_n \overset{p}{\to} X \Rightarrow g(X_n) \overset{p}{\to} g(X)$$
$$X_n \overset{a.s.}{\to} X \Rightarrow g(X_n) \overset{a.s.}{\to} g(X)$$
$$X_n \overset{d}{\to} X \Rightarrow g(X_n) \overset{d}{\to} g(X)$$

For the heuristics of the third line: If g is continuous, then $f \circ g$ is continuous and bounded for any continuous bounded f. Thus,

$$\mathbb{E}[f(g(X_n))] \to \mathbb{E}[f(g(x))]$$

Another important theorem we will need is Slutsky's Theorem.

Theorem 2. *(Slutsky's Theorem)*

(1) If c is constant, then

$$X_n \xrightarrow{d} c \Leftrightarrow X_n \xrightarrow{p} c$$

(2) If $X_n \xrightarrow{d} X$, $d(X_n, Y_n) \xrightarrow{p} 0$, then

$$Y_n \xrightarrow{d} X$$

(3) If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c$, then

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

The Slutsky's theorem allows us to ignore low order terms in convergence. Also, the following example shows that stronger impliations over part (3) may not be true.

Example 3: If $X_n \xrightarrow{d} \mathsf{N}(0, I)$, then $-X_n \xrightarrow{d} \mathsf{N}(0, I)$.

However,

$$\begin{pmatrix} X_n \\ -X_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ -Z \end{pmatrix}$$

where $Z \sim \mathsf{N}(0, I)$ instead of $\mathsf{N}(0, I)$. ♣

Sketch of Proof

(1) The " $\Leftarrow$ " direction is trivial and given in the previous sections. For " $\Rightarrow$ " direction of the theorem, take

$$f(x) = ||x - c|| \wedge 1 = \min\{||x - c||, 1\}$$

then

$$\mathbb{E}[||x_n - c|| \wedge 1] \to 0$$

(2) Let f be 1-Lipschitz and bounded by 1, then we have

$$\mathbb{E}[f(Y_n)] \in \mathbb{E}[f(X_n)] \pm \mathbb{E}[d(X_n, Y_n) \wedge 1]$$

Since $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ and $\mathbb{E}[d(X_n, Y_n) \wedge 1] \to 0$, we have

$$\mathbb{E}[f(Y_n)] \to \mathbb{E}[f(X)]$$

and thus $Y_n \to X$.

(3) We have

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} - \begin{pmatrix} X \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ Y_n - c \end{pmatrix} \xrightarrow{p} 0$$

By part (2),

$$\begin{pmatrix} X_n \\ c \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix} \Rightarrow \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

□

Consequences of Slutsky's Theorem:

If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$, then

$$X_n + Y_n \xrightarrow{d} X + c$$

$$Y_n X_n \xrightarrow{d} cX$$

If $c \neq 0$,

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$$

Proof Apply Continuous Mapping Theorem and Slutsky's Theorem and the statements can be proved. □

Note: For the third line of convergence, if $c \in \mathbb{R}^{d\times d}$ is a matrix, then (2) still holds. Moreover, if $\det(c) \neq 0$, (3) holds but

$$Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$$

because $c \to c^{-1}$ is continuous when $\det(c) \neq 0$.

Example 4: (t-type statistics:) Let $X_i \overset{\text{iid}}{\sim} P$, $\text{Cov}(X_i) = \Gamma > 0$, define

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_n)(X_i - \mu_n)^T$$

$$T_n = \frac{1}{\sqrt{n}} S_n^{-\frac{1}{2}} \sum_{i=1}^{n}(X_i - \mu)$$

Then $T_n \xrightarrow{d} \mathsf{N}(0, I)$.

The reason is that

$$\mu_n \xrightarrow{p} \mathbb{E}[X]$$

and

$$S_n \xrightarrow{p} \Gamma$$

Apply Slutsky's Theorem,

$$T_n - \frac{1}{\sqrt{n}}\Gamma^{-\frac{1}{2}} \sum_{i=1}^{n}(X_i - \mu) \xrightarrow{p} 0$$

♣

Big-O Notation: (in probability)

In this part we introduce the big-o and little-o notation in probbility.

Let X_n be random vectors, R_n be random variables. We say that $X_n = o_p(R_n)$ if $\exists$ random vectors Y_n such that

$$X_n = Y_n R_n$$
$$Y_n \xrightarrow{p} 0$$

This is called "little o-pea".

We say that $X_n = O_p(R_n)$ if $\exists$ random vectors Y_n where $Y_n = O_p(1)$. $Y_n = O_p(1)$ means that means $\{Y_n\}$ is uniformly tight. i.e.

$$\lim_{M\to\infty} \sup \sup_n \mathbb{P}(||Y_n|| \geq M) = 0$$

or $\forall \epsilon > 0$, $\exists$ M such that

$$\mathbb{P}(||Y_n|| \geq M) \leq \epsilon, \forall\, n$$

Comsequences:

With the definition above, we can get the following properties and lemma.

$$o_p(1) + o_p(1) = o_p(1)$$
$$O_p(1) + o_p(1) = O_p(1)$$
$$O_p(1) + O_p(1) = O_p(1)$$

Lemma 3. *Let function $R : \mathbb{R}^d \to \mathbb{R}^k$, with $R(0) = 0$. Let $X_n \xrightarrow{p} 0$, then*

(1) If $R(h) = o(||h||^p)$ as $h \to 0$, then

$$R(X_n) = o_p(||X_n||^p)$$

(2) If $R(h) = O(||h||^p)$ as $h \to 0$, then

$$R(X_n) = O_p(||X_n||^p)$$

Proof Define

$$g(h) = \begin{cases} \dfrac{R(h)}{||h||^p}, if\ h \neq 0 \\ 0, if\ h = 0 \end{cases}$$

(1) Then $g(h) \to 0$ as $h \to 0$. Thus, g is continuous at 0 and $X_n \xrightarrow{p} 0$. Apply Continuous Mapping Theorem(CMT), we get

$$g(X_n) \xrightarrow{p} 0$$

(2) $\exists$ M, $\delta > 0$ such that $||g(h)|| \leq M$, $\forall ||h|| \leq \delta$. Then

$$\Phi(||g(X_n)|| > M) \leq \mathbb{P}(||X_n|| \geq \delta) \to 0$$

so

$$g(X_n) = O_p(1)$$

$\square$