THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 1: Introduction

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

www.cs.cmu.edu/∼psarkar/teaching

Manegerial Stuff

- Instructor- Purnamrita Sarkar
- Course material and homeworks will be posted under `www.cs.cmu.edu/~psarkar/teaching/sds384.html`
- Office hours: Tuesdays 11-12pm. GDC —
- TA: TBD
- Homeworks are due Biweekly after class on thursdays
- Grading - 5 homeworks (60%), Midterm (20%), Final Project (20%)
- Books
  - Asymptotic Statistics, Aad van der Vaart. Cambridge. 1998.
  - Convergence of Stochastic Processes, David Pollard. Springer. 1984. Available on-line at http://www.stat.yale.edu/pollard/1984book/

Why do theory?

- Say you have estimated $\hat{\theta}_n$ from data $X_1, \ldots, X_n$. How do we know we have a "good" estimation method?
    - Does $\hat{\theta}_n \to \theta$? This brings us to **Stochastic Convergence**.
- How do I know if one estimation method is better than another?
    - Does the estimate from one converge faster than the other?
    - Does one algorithm work under broader parameter regimes, or weaker assumptions?
    - What is the optimal rate for a given estimation problem?

This class

- Consistency of parameter estimates
    - Stochastic Convergence
    - Concentration inequalities
    - Asymptotic normality of estimators
- Empirical processes, VC classes, covering numbers
- Asymptotic testing
- Examples of network clustering with a bit of random matrix theory

Stochastic Convergence

Assume that $X_n, n \geq 1$ and X are elements of a separable metric space (S, d).

Definition (Weak Convergence)

A sequence of random variable s converge in "law" or in "distribution" to a random variable X, i.e. $X_n \xrightarrow{d} X$ if $P(X_n \leq x) \to P(X \leq x)$ $\forall x$ at which $P(X \leq x)$ is continuous.

Definition (Convergence in Probability)

A sequence of random variables converge in "probability" to a random variable X, i.e. $X_n \xrightarrow{P} X$ if $\forall \epsilon > 0$, $P(d(X_n, X) \geq \epsilon) \to 0$.

Stochastic Convergence

Assume that $X_n, n \geq 1$ and X are elements of a separable metric space (S, d).

Definition (Almost Sure Convergence)

A sequence of random variables converge almost surely to a random variable X, i.e. $X_n \stackrel{a.s.}{\to} X$ if $P\left(\lim_{n\to\infty} d(X_n, X) = 0\right) = 1$.

Definition (Convergence in quadratic mean)

A sequence of random variables converge in quadratic mean to a random variable X, i.e. $X_n \stackrel{q.m}{\to} X$ if $E\left[d(X_n, X)^2\right] \to 0$.

Stochastic Convergence

Theorem

$$X_n \overset{a.s.}{\to} X \ , X_n \overset{q.m.}{\to} X \Rightarrow X_n \overset{P}{\to} X \Rightarrow X_n \overset{d}{\to} X$$

$$X_n \overset{d}{\to} c \Rightarrow X_n \overset{P}{\to} c$$

Continuous Mapping Theorem

Theorem

Let g be continuous on a set C where $P(X \in C) = 1$. Then,

$$X_n \stackrel{d}{\to} X \Rightarrow g(X_n) \stackrel{d}{\to} g(X)$$
$$X_n \stackrel{P}{\to} X \Rightarrow g(X_n) \stackrel{P}{\to} g(X)$$
$$X_n \stackrel{a.s.}{\to} X \Rightarrow g(X_n) \stackrel{a.s.}{\to} g(X)$$

Example

Let $X_n \xrightarrow{d} X$ where $X \sim N(0,1)$. Then $X_n^2 \xrightarrow{d} ?$

Example

Let $X_n \xrightarrow{d} X$ where $X \sim N(0,1)$. Then $X_n^2 \xrightarrow{d}$?

- Use $g(x) = x^2$.

Example

Let $X_n \overset{d}{\to} X$ where $X \sim N(0,1)$. Then $X_n^2 \overset{d}{\to}?$

- Use $g(x) = x^2$.
- Use $X^2 \sim \chi_1^2$.

Example

Let $X_n \xrightarrow{d} X$ where $X \sim N(0,1)$. Then $X_n^2 \xrightarrow{d} ?$

- Use $g(x) = x^2$.
- Use $X^2 \sim \chi_1^2$.
- So $X_n^2 \xrightarrow{d} \chi_1^2$

Example-continuity points

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. We have $\bar{X}_n - \mu \xrightarrow{d} 0$. Consider $g(x) = 1_{x>0}$. Then $g((\bar{X}_n - \mu)^2) \xrightarrow{d} ?$

Example-continuity points

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. We have $\bar{X}_n - \mu \xrightarrow{d} 0$. Consider $g(x) = 1_{x>0}$. Then $g((\bar{X}_n - \mu)^2) \xrightarrow{d} ?$

- Using Continuous Mapping Theorem, $(\bar{X}_n - \mu)^2 \xrightarrow{d} 0$

Example-continuity points

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. We have $\bar{X}_n - \mu \xrightarrow{d} 0$. Consider $g(x) = 1_{x>0}$. Then $g((\bar{X}_n - \mu)^2) \xrightarrow{d} ?$

- Using Continuous Mapping Theorem, $(\bar{X}_n - \mu)^2 \xrightarrow{d} 0$
- Can we use Continuous Mapping Theorem to claim that $g(\bar{X}_n - \mu)^2 \xrightarrow{d} 0$?

Example-continuity points

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. We have $\bar{X}_n - \mu \xrightarrow{d} 0$. Consider $g(x) = 1_{x>0}$. Then $g((\bar{X}_n - \mu)^2) \xrightarrow{d} ?$

- Using Continuous Mapping Theorem, $(\bar{X}_n - \mu)^2 \xrightarrow{d} 0$
- Can we use Continuous Mapping Theorem to claim that $g(\bar{X}_n - \mu)^2 \xrightarrow{d} 0$?
- NO. Because, 0 is a random variable whose mass is at 0, where g is discontinuous.

Portmanteau Theorem

Theorem

The following are equivalent.

- $X_n \xrightarrow{d} X$
- $E[f(X_n)] \to E[f(X)]$ *for all bounded and continuous* f*.*
- $E[f(X_n)] \to E[f(X)]$ *for all bounded and Lipschitz* f*.*
- $E[e^{it^T X_n}] \to E[e^{it^T X_n}]$, $\forall t \in \mathbb{R}^k$*. (Levy's continuity theorem)*
- $t^T X_n \xrightarrow{d} t^T X \ \forall t \in \mathbb{R}^k$*. (Cramer-Wold device)*
- $\liminf_n E[f(X_n)] \geq E[f(X)]$ *for all non-negative continuous* f
- $\limsup_n P(X_n \in F) \leq P(X \in F)$ *for all closed* F
- $\liminf_n P(X_n \in F) \geq P(X \in F)$ *for all open* F
- $P(X_n \in B) \to P(X \in B)$ *for all continuity sets* B *(*$P(X \in \partial B) = 0$*)*

Example-bounded

Consider $f(x) = x$ and

$$X_n = \begin{cases} n & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases}$$

Example-bounded

Consider $f(x) = x$ and

$$X_n = \begin{cases} n & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases}$$

- $X_n \xrightarrow{d} 0$, but $E[X_n] \to ?$

Example-bounded

Consider $f(x) = x$ and

$$X_n = \begin{cases} n & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases}$$

- $X_n \xrightarrow{d} 0$, but $E[X_n] \to ?$
- $E[X_n] = 1$. What went wrong?

Example-bounded

Consider $f(x) = x$ and

$$X_n = \begin{cases} n & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases}$$

- $X_n \xrightarrow{d} 0$, but $E[X_n] \to ?$
- $E[X_n] = 1$. What went wrong?
- $f(x)$ is not bounded.

Putting everything together

Theorem

$$X_n \xrightarrow{d} X \text{ and } d(X_n, Y_n) \xrightarrow{P} 0 \Rightarrow Y_n \xrightarrow{d} X \tag{1}$$

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} c \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, c) \tag{2}$$

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \Rightarrow (X_n, Y_n) \xrightarrow{P} (X, Y) \tag{3}$$

Putting everything together

Theorem

$$X_n \xrightarrow{d} X \textit{ and } d(X_n, Y_n) \xrightarrow{P} 0 \Rightarrow Y_n \xrightarrow{d} X \tag{1}$$

$$X_n \xrightarrow{d} X \textit{ and } Y_n \xrightarrow{d} c \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, c) \tag{2}$$

$$X_n \xrightarrow{P} X \textit{ and } Y_n \xrightarrow{P} Y \Rightarrow (X_n, Y_n) \xrightarrow{P} (X, Y) \tag{3}$$

- Eq 3 does not hold if we replace convergence in probability by convergence in distribution.

Putting everything together

Theorem

$$X_n \xrightarrow{d} X \text{ and } d(X_n, Y_n) \xrightarrow{P} 0 \Rightarrow Y_n \xrightarrow{d} X \tag{1}$$

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} c \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, c) \tag{2}$$

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \Rightarrow (X_n, Y_n) \xrightarrow{P} (X, Y) \tag{3}$$

- Eq 3 does not hold if we replace convergence in probability by convergence in distribution.
- Example: $X_n \sim N(0,1), Y_n = -X_n$. $X \perp Y$ and X, Y are independent standard normal random variables.

Putting everything together

Theorem

$$X_n \xrightarrow{d} X \text{ and } d(X_n, Y_n) \xrightarrow{P} 0 \Rightarrow Y_n \xrightarrow{d} X \tag{1}$$

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} c \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, c) \tag{2}$$

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \Rightarrow (X_n, Y_n) \xrightarrow{P} (X, Y) \tag{3}$$

- Eq 3 does not hold if we replace convergence in probability by convergence in distribution.
- Example: $X_n \sim N(0,1), Y_n = -X_n$. $X \perp Y$ and X, Y are independent standard normal random variables.
- Then $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. But $(X_n, Y_n) \xrightarrow{d} (X, -X)$, not $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

Putting everything together

Theorem (Slutsky's theorem)

$X_n \stackrel{d}{\to} X$ and $Y_n \stackrel{d}{\to} c$ imply that

$$X_n + Y_n \stackrel{d}{\to} X + c$$

$$X_n Y_n \stackrel{d}{\to} cX$$

$$X_n / Y_n \stackrel{d}{\to} X / c$$

Putting everything together

Theorem (Slutsky's theorem)

$X_n \stackrel{d}{\to} X$ and $Y_n \stackrel{d}{\to} c$ imply that

$$X_n + Y_n \stackrel{d}{\to} X + c$$

$$X_n Y_n \stackrel{d}{\to} cX$$

$$X_n / Y_n \stackrel{d}{\to} X / c$$

- *Does $X_n \stackrel{d}{\to} X$ and $Y_n \stackrel{d}{\to} Y$ imply $X_n + Y_n \stackrel{d}{\to} X + Y$?*

Putting everything together

Theorem (Slutsky's theorem)

$X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ imply that

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

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

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

- *Does $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ imply $X_n + Y_n \xrightarrow{d} X + Y$?*
- *Take $Y_n = -X_n$, and X, Y as independent standard normal random variables. $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ but $X_n + Y_n \xrightarrow{d} 0$.*

Using all this

If $X_1, \ldots X_n$ are i.i.d. random variables with mean μ and variance σ^2, prove that $\sqrt{n}\dfrac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$.

Using all this

If $X_1, \dots X_n$ are i.i.d. random variables with mean μ and variance σ^2, prove that $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$.

- First note that $S_n = \frac{1}{n-1}\sum_i X_i^2 - \bar{X}_n^2 = \frac{n}{n-1}\frac{\sum_i X_i^2}{n} - \bar{X}_n^2$

Using all this

If $X_1, \ldots X_n$ are i.i.d. random variables with mean μ and variance σ^2, prove that $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$.

- First note that $S_n = \frac{1}{n-1}\sum_i X_i^2 - \bar{X}_n^2 = \frac{n}{n-1}\frac{\sum_i X_i^2}{n} - \bar{X}_n^2$
- Law of large numbers give $\frac{\sum_i X_i^2}{n} \xrightarrow{P} E[X^2]$ and $X_n \xrightarrow{P} \mu$.

Using all this

If $X_1, \ldots X_n$ are i.i.d. random variables with mean μ and variance σ^2, prove that $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$.

- First note that $S_n = \frac{1}{n-1}\sum_i X_i^2 - \bar{X}_n^2 = \frac{n}{n-1}\frac{\sum_i X_i^2}{n} - \bar{X}_n^2$
- Law of large numbers give $\frac{\sum_i X_i^2}{n} \xrightarrow{P} E[X^2]$ and $X_n \xrightarrow{P} \mu$.
- So $(\frac{\sum_i X_i^2}{n}, X_n) \xrightarrow{P} (E[X^2], \mu)$ and now using the continuous mapping theorem, $S_n^2 \xrightarrow{P} \sigma^2$.

Using all this

If $X_1, \ldots X_n$ are i.i.d. random variables with mean μ and variance σ^2, prove that $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$.

- First note that $S_n = \frac{1}{n-1}\sum_i X_i^2 - \bar{X}_n^2 = \frac{n}{n-1}\frac{\sum_i X_i^2}{n} - \bar{X}_n^2$
- Law of large numbers give $\frac{\sum_i X_i^2}{n} \xrightarrow{P} E[X^2]$ and $X_n \xrightarrow{P} \mu$.
- So $(\frac{\sum_i X_i^2}{n}, X_n) \xrightarrow{P} (E[X^2], \mu)$ and now using the continuous mapping theorem, $S_n^2 \xrightarrow{P} \sigma^2$.
- Finally, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ using CLT.

Using all this

If $X_1, \dots X_n$ are i.i.d. random variables with mean μ and variance σ^2, prove that $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$.

- First note that $S_n = \frac{1}{n-1}\sum_i X_i^2 - \bar{X}_n^2 = \frac{n}{n-1}\frac{\sum_i X_i^2}{n} - \bar{X}_n^2$
- Law of large numbers give $\frac{\sum_i X_i^2}{n} \xrightarrow{P} E[X^2]$ and $X_n \xrightarrow{P} \mu$.
- So $(\frac{\sum_i X_i^2}{n}, X_n) \xrightarrow{P} (E[X^2], \mu)$ and now using the continuous mapping theorem, $S_n^2 \xrightarrow{P} \sigma^2$.
- Finally, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ using CLT.
- Now using Slutsky's lemma, $\sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{d} N(0,1)$ using CLT.

Uniformly tight

Definition

X is defined to be "tight" if $\forall \epsilon > 0\ \exists M$ for which,

$$P(\|X\| > M) < \epsilon$$

$\{X_n\}$ is defined to uniformly tight if $\forall \epsilon > 0\ \exists M$ for which,

$$\sum_n P(\|X_n\| > M) < \epsilon$$

Prohorov's theorem

Theorem

- $X_n \xrightarrow{d} X \Rightarrow \{X_n\}$ *is UI.*
- $\{X_n\}$ *is UI implies that, there exists a subsequence* $\{n_j\}$ *such that* $X_{n_j} \xrightarrow{d} X$.

Notation for rates, small oh-pee and big oh-pee

Definition

- The small o_P:

$$X_n = o_P(1) \Leftrightarrow X_n \xrightarrow{P} 0$$
$$X_n = o_P(R_n) \Leftrightarrow X_n = Y_n R_n \text{ and } Y_n = o_P(1)$$

  X_n is vanishing in probability

- The big O_P:

$$X_n = O_P(1) \Leftrightarrow \{X_n\} \text{ is UI}$$
$$X_n = O_P(R_n) \Leftrightarrow X_n = Y_n R_n \text{ and } Y_n = O_P(1)$$

  X_n lies within a ball of finite radius with high probability

How do they interact

$$o_P(1) + o_P(1) = o_P(1).$$
$$o_P(1) + O_P(1) = O_P(1).$$
$$O_P(1)o_P(1) = o_P(1).$$
$$1 + O_P(1) = O_P(1).$$
$$(1 + o_P(1))^{-1} = O_P(1).$$
$$o_P(O_P(1)) = o_P(1).$$

$$X_n \xrightarrow{P} 0, R(h) = o(\|h\|^p) \Rightarrow R(X_n) = o_P(\|X_n\|^p)$$
$$X_n \xrightarrow{P} 0, R(h) = O(\|h\|^p) \Rightarrow R(X_n) = O_P(\|X_n\|^p)$$

Be careful:

$$e^{o_P(1)} \neq o_P(1)$$

$O_P(1) + O_P(1)$ Can actually be $o_P(1)$ because of cancellation.