Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 1

Yun Yang

- Course information.
- Brief review.

Instructor and TA

- Yun Yang, Assistant Professor, Department of Statistics, OSB 209G
- yyang@stat.fsu.edu
- Office hour: Tuesday & Thursday 1:15pm – 2:00pm or by appointment (at least 24 hours in advance)
- TA: Shuang Zhou (shuang.zhou@stat.fsu.edu)
- TA Office Hour: Monday 3:30pm – 4:30pm, Wednesday: 10:30am – 11:30am, 448 Dirac Science Library, or by appointment (at least 24 hours in advance)

Course website

- Canvas: https://fsu.instructure.com/
- Syllabus
- Announcements (*mostly via email*)
- Lecture Notes and Slides
- Homeworks and Solutions
- Grades
- Others

Grades

- Homework: 40% (best 5 out of total 6 assignments)
- Attendance: 5%
- Midterm: 25% (February 27 in class)
- Final project: 30% (proposal due March 15; report due April 27)

Check the syllabus online for details.

Homework

- Roughly biweekly.
- Due at the beginning of classes.
- Late homework will not be accepted.
- Reinforces lecture and examples
- Prepares you for exams.
- **You will get at least one week for each assignment. Don't wait until the due date!**

Textbooks

No required textbooks, recommended texts:

- *Asymptotic Statistics*, Aad van der Vaart, Cambridge, 1998.
- *Weak convergence and empirical processes: with applications to statistics*, Aad van der Vaart and Jon Wellner, Springer, 2000.
- *Empirical processes in M-estimation*, Sara van de Geer, Cambridge University Press, 2009.

Prerequisite

- Distribution Theory (STA 5326)
- Statistical Inference (STA 5327)
- Advanced Probability and Inference I (STA 6346).

Talk to me if you are unsure of your background.

Goal

Provide theoretical background and mathematical tools for the finite sample analysis of modern data science methods.

Example

We have a sample of size n from a density p_θ, where parameter $\theta \in \mathbb{R}^d$. Some statistical procedure gives estimator $\widehat{\theta}_n$.

- Consistent? i.e. $\widehat{\theta}_n \to \theta^*$?
- Rate of convergence? Is it optimal?

Finite sample analysis

Find a function $f(n, d, \delta)$, such that: for any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that

$$\ell(\widehat{\theta}_n, \theta^*) \leq f(n, d, \delta),$$

where ℓ is a loss function such as the squared loss $\ell(x, x') = \|x - x'\|^2$.

Finite Sample Analysis: Why?

Example (Linear regression with increasing dimension)

Linear model: for $i = 1, 2, \ldots, n$,

$$Y_i = X_i^T \theta + w_i, \quad w_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Here $\theta \in \mathbb{R}^d$ is the unknown regression coefficient vector.

Least square estimator: $\widehat{\theta}_n \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \theta)^2.$

- For fixed d, we have $\|\widehat{\theta}_n - \theta^*\|^2 = O_P(n^{-1})$. How about $d \gg 1$?
- One solution is to consider the "double asymptotics"

Finite Sample Analysis: Why?

Example (Central limit theorem in increasing dimension)

$X_1, X_2, \ldots, X_n$ are i.i.d. d-dim random vectors satisfying

$$\mathbb{E}[X_i] = 0 \quad \text{and} \quad \mathrm{Cov}[X_i] = I_d, \quad i = 1, \ldots, n.$$

When $d \gg 1$, do we still have

$$\text{distribution of } Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \text{ close to } \mathcal{N}(0, I_d)?$$

- Related: limiting distribution of maximum likelihood estimator (MLE)
- Levy's convergence theorem does not apply. Need other techniques.

Possible Course Flow

- Concentration inequalities
- Uniform laws of large number
- Metric entropy, chaining, asymptotic equicontinuity
- High-dimensional regression
- High-dimensional covariance matrix estimation
- Non-parametric regression
- Minimax lower bounds

Since this is a very diverse classroom, we will adjust the course flow to fit most people's background.

Different types of convergence

$X, X_1, X_2, \ldots$ are random vectors, d is a distance

Definition

X_n **converges almost surely** to X (write $X_n \overset{a.s}{\to} X$) means $d(X_n, X) \to 0$ a.s.

Definition

X_n **converges in probability** to X (write $X_n \overset{P}{\to} X$) means that, for any $\varepsilon > 0$, $\mathbb{P}(d(X_n, X) \geq \varepsilon) \to 0$.

Definition

X_n **converges in distribution** (or **weakly converges**) to X (write $X_n \rightsquigarrow X$) means their distribution functions satisfy $F_n(x) \to F(x)$ at all continuity points of F.

Relations

Theorem

$$X_n \xrightarrow{a.s} X \Longrightarrow X_n \xrightarrow{P} X \Longrightarrow X_n \rightsquigarrow X$$

$$X_n \xrightarrow{P} c \Longleftrightarrow X_n \rightsquigarrow c.$$

Note: For $X_n \xrightarrow{a.s} X$ and $X_n \xrightarrow{P} X$, X_n and X must be on the sample space of the same probability space. But not convergence in distribution.

Convergence in distribution

Theorem (Portmanteau)

The following statements are equivalent.

a. $X_n \rightsquigarrow X$.

b. $\liminf \mathbb{P}(X_n \in U) \geq \mathbb{P}(X \in U)$, *for all open* U.

c. $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$, *for all closed* F.

d. $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$, *for all continuity sets* A
(i.e. $\mathbb{P}(X \in \delta A) = 0$, *where* δA *denotes the boundary of* A*).*

e. $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ *for all bounded and continuous function* f.

f. $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ *for all bounded and Lipschitz function* f.

Why continuous? Consider $f(x) = I(x > 0)$ and $X_n = 1/n$;
Why boundedness? Consider $f(x) = x$ and

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

Continuous mapping theorem

Suppose $f : \mathbb{R}^d \to \mathbb{R}^k$ is a continuous function.

Theorem

$$X_n \rightsquigarrow X \Longrightarrow f(X_n) \rightsquigarrow f(X),$$
$$X_n \xrightarrow{P} X \Longrightarrow f(X_n) \xrightarrow{P} f(X),$$
$$X_n \xrightarrow{a.s} X \Longrightarrow f(X_n) \xrightarrow{a.s} f(X).$$

Slutsky's lemma

Theorem

$X_n \rightsquigarrow X$ *and* $Y_n \rightsquigarrow c$ *imply*

$$X_n + Y_n \rightsquigarrow X + c,$$
$$X_n Y_n \rightsquigarrow cX,$$
$$(c \neq 0) \quad Y_n^{-1} X_n \rightsquigarrow c^{-1} X.$$

Small O and big O notation

Definition

$$X_n = o_P(1) \Longrightarrow X_n \xrightarrow{P} 0;$$

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

$$X_n = O_P(1) \Longrightarrow X_n \text{ uniformly tight: for any } \varepsilon > 0,$$
$$\text{there is an } M \text{ such that } \sup_n \mathbb{P}(|X_n| > M) < \varepsilon;$$

$$X_n = O_P(R_n) \Longrightarrow X_n = Y_n R_n \text{ and } Y_n = O_P(1).$$

o_P and O_P specify rates of convergence. $o_P(R_n)$ means strictly slower than R_n. $O_P(R_n)$ means within constant the same as R_n.

Relations

$$
\begin{aligned}
o_P(1) + o_P(1) &= o_P(1), \\
o_P(1) + O_P(1) &= O_P(1), \\
o_P(1) O_P(1) &= o_P(1), \\
(1 + o_P(1))^{-1} &= O_P(1), \\
o_P(O_P(1)) &= o_P(1).
\end{aligned}
$$

Moments via tails

Let X be a random variable and $p \in (0, \infty)$.

$$\mathbb{E}|X|^p = \int_0^\infty p\, t^{p-1}\, \mathbb{P}(|X| \geq t)\, dt.$$

In particular, if X is a nonnegative random variable, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt.$$

Example

If an estimator $\widehat{\theta}_n$ satisfies

$$\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| > t) \leq \exp\{-n\, t^2/2\},$$

then

$$\mathbb{E}[\|\widehat{\theta}_n - \theta^*\|] \leq \int_0^\infty \exp\{-n\, t^2/2\}\, dt \leq \frac{C}{\sqrt{n}}.$$

Tails via moments

Markov's inequality

For any nonnegative random variable X and $t > 0$, we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Chebyshev's inequality

For any random variable X, $k > 0$ and $t > 0$, we have

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X - \mu|^k]}{t^k}.$$

where $\mu = \mathbb{E}[X]$.