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

SDS 384 11: Theoretical Statistics

Lecture 19: Overview

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

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. $X_n \overset{a.s.}{\rightarrow} X$ if $P\left(\lim_{n\rightarrow\infty} d(X_n, X) = 0\right) = 1$.

- If you think about a (scalar) random variable as a function that maps events to a real number, almost sure convergence means $P(\omega \in \Omega : \lim_{n\rightarrow\infty} X_n(\omega) = X(\omega)) = 1$

Definition (Convergence in quadratic mean)

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

Continuous Mapping Theorem

Theorem

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

$$X_n \stackrel{d}{\to} X \Rightarrow g(X_n) \stackrel{d}{\to} g(X)$$

$$X_n \stackrel{P}{\to} X \Rightarrow g(X_n) \stackrel{P}{\to} g(X)$$

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

- What about continuous mapping with quadratic mean?

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)

$X_n \xrightarrow{d} X$ 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 $X_n + Y_n \xrightarrow{d} 0$.*

Lindeberg-feller CLT for triangular arrays

Theorem (Ordinary Central limit theorem)

$X_1, \ldots, X_n \overset{iid}{\sim} f$ *with* $E|X_i| \leq \infty$, $E[X_1] = 0$. *If* $E[X_i^2] = \sigma^2$, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

$$X_{11}$$
$$X_{21}, X_{22}$$
$$X_{21}, X_{22}, X_{23}$$
$$\ldots$$

Theorem (Lindeberg-feller)

For each n let $(X_{ni})_{i=1}^n$ *be independent random variables with mean zero and variance* σ_{ni}^2. *Let* $Z_n = \sum_{i=1}^n X_{ni}$ *and* $B_n^2 = var(Z_n)$. *Then* $Z_n/B_n \xrightarrow{d} N(0,1)$, *as long as the* **Lindeberg condition** *holds.*

The Lindeberg condition

Definition (Lindeberg condition)

For every $\epsilon > 0$,

$$\frac{1}{B_n^2}\sum_{j=1}^{n} E[X_{nj}^2 1(|X_{nj}| \geq \epsilon B_n)] \to 0 \text{ as } n \to \infty \tag{4}$$

Converse: If $\frac{\sigma_{nj}^2}{B_n^2} \to 0$ as $n \to \infty$, i.e. no one variance plays a significant role in the limit, and if $Z_n/B_n \xrightarrow{d} N(0,1)$, then the Lindeberg condition holds.

Necessary and Sufficient: If $\frac{\sigma_{nj}^2}{B_n^2} \to 0$, the the Lindeberg condition is necessary and sufficient to show the CLT.

Example

Let $X_1, \dots, X_n$ be independent random variables with mean zero and variance one. Do you think $\sqrt{n}\bar{X}_n \xrightarrow{d} N(0,1)$?

- $$X_{nj} = \begin{cases} 2j & \text{w.p. } \dfrac{1}{8j^2} \\ 0 & \text{w.p. } 1 - \dfrac{1}{4j^2} \\ -2j & \text{w.p. } \dfrac{1}{8j^2} \end{cases}$$

- $E[X_{nj}] = 0$ and $\text{var}(X_{nj}) = 1$. $B_n^2 = n$.
- Lets check the Lindeberg condition with $\epsilon = 1$.

$$\frac{1}{n}\sum_j E[X_{nj}^2 1(|X_{nj}| \geq \sqrt{n})] = \frac{1}{n}\sum_j 2 \times 4j^2 1(2j \geq \sqrt{n})\frac{1}{8j^2} = \frac{1}{n}\sum_{j \geq \sqrt{n}/2} 1 \to 1$$

- Since $\sigma_{nj}^2 / B_n^2 = 1/n \to 0$, this implies that the CLT does not hold for the sum.

Chernoff bound

- We have done CLT, but it does not give us explicit tail bounds.
- Lets look at concentration inequalities.

Theorem (Chernoff bound for Bernoullis)

Let $X_i \in \{0,1\}$ *be independent random variables with* $E[X_i] = p_i$. *Let* $X := \sum_i X_i, \mu := \sum_i p_i$. *For* $0 < \delta < 1$,

$$P(X \geq \mu(1+\delta)) \leq e^{-\delta^2 \mu/3} \qquad P(X \leq \mu(1-\delta)) \leq e^{-\delta^2 \mu/2}$$

- How about subgaussian r.v.s?

Sub-Gaussian random variables

Theorem

For $X_1, \dots, X_n$ independent sub-gaussian random variables with sub-gaussian parameters σ_i^2 and $E[X_i] = \mu_i$, for $\forall t > 0$,

$$P\left(\sum_i (X_i - \mu_i) \geq t\right) \leq e^{-\frac{t^2}{2\sum_i \sigma_i^2}}$$

- If $X_i \in [a, b]$, $E[X_i] = 0$, using Hoeffding's lemma we get:

$$P\left(\sum_i X_i \geq t\right) \leq e^{-\frac{2t^2}{n(b-a)^2}}$$

- If $X_i \sim N(0, \sigma^2)$, we immediately get back the chernoff bound for Gaussians.

Sub-exponential random variables

Definition

X is sub-exponential with parameters (ν, b) if, $\forall |\lambda| < 1/b$,

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \nu^2}{2}$$

Concentration

Theorem

Let X be a sub-exponential random variable with parameters (ν, b). Then,

$$P(X \geq \mu + t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{b} \\ e^{-\frac{t}{2b}} & \text{if } t \geq \frac{\nu^2}{b} \end{cases}$$

- For small t this is sub-gaussian in nature, whereas for large t the exponent decays linearly with t.

Bernstein's condition and the sub-exponential property

Definition

A random variable with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, if $|E[(X-\mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2}$ for $k \geq 2$.

Theorem

If X ($E[X] = \mu, var(X) = \sigma^2$) satisfies the Bernstein condition with parameter $b > 0$, then X is sub-exponential with $(\sqrt{2}\sigma, 2b)$.

Theorem

If X with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, then

$$P(|X-\mu| \geq t) \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \tag{5}$$

How about martingale inequalities?

Theorem

Let $f : \mathcal{X}^n \to \mathcal{R}$ *satisfy the following bounded difference condition* $\forall x_1, \ldots, x_n, x_i' \in \mathcal{X}$*:*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq B_i,$$

then, $P(|f(X) - E[f(X)]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right)$

- Note that this boils down to Hoeffding's when f is the sum of bounded random variables.

Recall-Lipschitz functions of Gaussian random variables

Theorem (LG:Lipschtiz functions of Gaussians)

Let $(X_1, \ldots, X_n)$ be a vector of iid $N(0,1)$ random variables. Let $f : \mathcal{R}^n \to \mathcal{R}$ be $L-$Lipschitz w.r.t the Euclidean norm. Then $f(X) - E[f(X)]$ is sub-gaussian with parameter at most L, i.e. $\forall t \geq 0$,

$$P\left(|f(X) - E[f(X)]| \geq t\right) \leq e^{-\frac{t^2}{2L^2}}$$

- So a $L-$Lipschitz function of n gaussian random variables behave like a subgaussian with variance proxy L^2.

Convex Lipschitz functions of bounded random variables

Theorem

Consider a convex function $f : \mathcal{R}^n \to \mathcal{R}$ with Lipschitz constant L. Also consider n iid random variables $X_1, \ldots, X_n \in \{-1, 1\}$. We have for $t > 0$

$$P(|f(X) - M_f| \geq t) \leq 4 \exp\left(-\frac{t^2}{16L^2}\right),$$

where M_f is the median of f.

- Often the median can be replaced by the mean with a little give in the t.

Efron Stein inequality

- Consider n independent random variables in some metric space $\mathcal{X}$.
- Consider a function $g : \mathcal{X}^n \to \mathbb{R}$
- Let $Z := g(X_1, \ldots, X_n)$
- We are interested in computing $\mathrm{var}(g(X_1, \ldots, X_n))$
- Define $E_i(Z) = E[Z|X_{1:i-1}, X_{i+1:n}]$

An upper bound

Theorem

$$var(Z) \leq \sum_{i=1}^{n} E\left[Z - E_i[Z]\right]^2 \leq \inf_{Z_i} E[Z - Z_i]^2.$$

where Z_i are measurable and square integrable functions of $X_1^n \setminus X_i$.

Theorem

Let $X_1', \ldots X_n'$ denote an independent copy of $X_1, \ldots, X_n$. Let $Z_i' = g(X_{1:i-1}, X_i', X_{i+1:n})$. We have:

$$var(Z) \leq \frac{1}{2} \sum_i E[(Z - Z_i')^2].$$

Example

This is one of the basic operations research problems. Given n numbers $x_1, \ldots, x_n \in [0,1]$, the question is the following: what is the minimal number of "bins" into which these numbers can be packed such that the sum of the numbers in each bin doesn't exceed one. Let $f(x_1, \ldots, x_n)$ be this minimum number. Show that $\mathrm{var}(f) \leq n/4$

- Now clearly by changing one of the x_i's, the value of $f(x_1, \ldots, x_n)$ cannot change by more than one.
- So taking

$$Z_i = (\sup_x f(x_1^{i-1}, x, x_{i+1}^n) + \inf_x f(x_1^{i-1}, x, x_{i+1}^n))/2$$

gives the answer.

Self bounding functions

Definition

A non-negative function $g : \mathcal{X}^n \to \mathcal{R}$ has the self bounding property if there exist functions $g_i : \mathcal{X}^{n-1} \to \mathcal{R}$ such that for all $x_1, \ldots, x_n \in \mathcal{X}$ and $i \in [n]$,

- $0 \leq g(x_1, \ldots, x_n) - g_i(x_{1:i-1}, x_{i+1:n}) \leq 1$
- $\sum_i \left(g(x_1, \ldots, x_n) - g_i(x_{1:i-1}, x_{i+1:n})\right) \leq g(x_1, \ldots, x_n)$

- Clearly, $\sum_i (g(x_{1:n}) - g_i(x_{1:i-1}, x_{i+1:n}))^2 \leq g(x_1, \ldots, x_n) =: Z$
- Now Theorem 1 gives:

$$\mathrm{var}(Z) \leq \sum_i E[(Z - E_i[Z])^2] \leq \sum_i E[(Z - g_i(x_{1:i-1}, x_{i+1:n}))^2] \leq E[g(x_{1:n})]$$

- So $\mathrm{var}(Z) \leq E[Z]$

Concentration of self bounding functions

Theorem

Consider $Z := g(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are independent random variables. For all $t \geq 0$,

$$P\left(Z \geq E[Z] + t\right) \leq \exp\left(-\frac{t^2}{2(EZ + t/3)}\right)$$

$$P\left(Z \leq E[Z] - t\right) \leq \exp\left(-\frac{t^2}{2EZ}\right)$$

Uniform laws and Rademacher complexity

- We can show that $\|\hat{P}_n - P\|_{\mathcal{F}} \leq 2\mathcal{R}_{\mathcal{F}} + \epsilon$ with prob. $1 - e^{-n\epsilon^2/2}$.
- $\mathcal{R}_{\mathcal{F}} = E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right|$ measures the maximum possible correlation (over all $f \in \mathcal{F}$) between the vector $(f(X_1), \ldots, f(X_n))$ and the "noise vector" $(\epsilon_1, \ldots, \epsilon_n)$.

Bounding $\mathcal{R}_{\mathcal{F}}$

Theorem

Let $A \subseteq \mathcal{R}^n$*,* $R = \max_{a \in A} \|a\|$*,*

$$E \sup_{a \in A} \langle \epsilon, a \rangle \leq \sqrt{2R^2 \log |A|}.$$

And,

$$E \sup_{a \in A} |\langle \epsilon, a \rangle| \leq \sqrt{2R^2 \log |2A|}.$$

This holds for general subgaussian RVs too.

Rademacher Complexity for binary function classes

$$\begin{aligned}
\|\hat{P}_n - P\|_{\mathcal{F}} &\leq 2\mathcal{R}_{\mathcal{F}} + \epsilon = 2E[E[\sup_{f\in\mathcal{F}} \sum_i \epsilon_i f(X_i)/n]|X] + \epsilon \\
&\leq 2E\sqrt{\frac{2\log(|\mathcal{F}(X_1^n) \cup -\mathcal{F}(X_1^n)|)}{n}} + \epsilon \\
&\leq \sqrt{\frac{8\log 2\max_X |\mathcal{F}(X_1^n)|}{n}} + \epsilon
\end{aligned}$$

- How do I control $|\mathcal{F}(X_1^n)|$?
- How big is $\max_X |\mathcal{F}(X_1^n)|$?

Growth function

Definition

For a binary valued function class $\mathcal{F}$, the growth function is:

$$\Pi_{\mathcal{F}}(n) = \max\{|\mathcal{F}(x_1^n)| x_1, \ldots, x_n \in \mathcal{X}\}$$

- $\mathcal{X}$ could be $\mathcal{R}^d$.
- $\mathcal{R}_{\mathcal{F}} \leq \sqrt{\frac{2\log(2\Pi_{\mathcal{F}}(n))}{n}}$
- $\Pi_{\mathcal{F}}(n) \leq 2^n$ (which is not really useful)
- We are looking for $\Pi_{\mathcal{F}}(n)$ growing polynomially with n.
- Using Sauer's lemma we know that $\Pi_{\mathcal{F}}(n) \leq (ne/d)^d$

VC dimension-to remember

Example

Let $\mathcal{F} = \{1_{(-\infty,t]} : t \in \mathcal{R}\}$ and $\mathcal{X} = \mathcal{R}$. Then $d_{VC}(\mathcal{F}) = 1$.

- First show that there exists some configuration of one point, which can be shattered by $\mathcal{F}$.
    - For any point x, if x has label 1, use $t > x$
    - If x has label 0, use $t < x$.
- Now show that there exists no two points which can be shattered by $\mathcal{F}$. (this takes a bit of an argument in more complex cases.)
    - For any two points (x, y) the labeling $(0, 1)$ cannot be achieved by any function in $\mathcal{F}$.

Moving away from binary functions

- You are interested in bounding $\mathcal{R}_\mathcal{F}$
- $\mathcal{R}_\mathcal{F}$ can be thought of as $E \sup_f |\langle \epsilon, f(X_1^n) \rangle| = E \sup_\theta |X_\theta|$, where X_θ is sub gaussian process wrt metric $\|f(X_1^n) - g(X_1^n)\|_2$, for $f, g \in \mathcal{F}$.
-

$$\begin{aligned}
\mathcal{R}_\mathcal{F} &= E \sup_\theta |E_{\theta_0}(X_\theta - X_{\theta_0})| \leq E \sup_\theta E_{\theta_0} |(X_\theta - X_{\theta_0})| \\
&\leq E \sup_\theta |E_{\theta_0}(X_\theta - X_{\theta_0})| \leq E \sup_\theta E_{\theta_0} |(X_\theta - X_{\theta_0})| \\
&\leq E \sup_{\theta, \theta'} (X_\theta - X_{\theta'})
\end{aligned}$$

Upper bound by 1 step discretization

Theorem

(1-step discretization bound). Let $\{X_\theta, \theta \in \mathcal{T}\}$ be a zero-mean sub-Gaussian process with respect to the metric d_X. Then for any $\delta > 0$, we have

$$E\left[\sup_{\theta,\theta' \in \mathcal{T}} (X_\theta - X_{\theta'})\right] \leq 2E\left[\sup_{\substack{\theta,\theta' \in \mathcal{T} \\ d_X(\theta,\theta') \leq \delta}} (X_\theta - X_{\theta'})\right] + 2D\sqrt{\log N(\delta; \mathcal{T}, d_X)},$$

where $D := \max_{\theta,\theta' \in \Theta} d_X(\theta, \theta')$.

- The mean zero condition gives us: $E[\sup_{a \in \mathcal{A}} a^T X] \leq E[\sup_{a,a' \in \mathcal{A}} (a^T X - a'^T X)]$
- $a^T X$ is sub Gaussian w.r.t the $\|.\|_2$ norm.
- $D = 2\mathcal{W}$.
- Then optimize. You will also need more information about $\mathcal{A}$ to make sure that you can calculate the covering number.

Putting everything in place

- First we do convergence, since it shows up everywhere.
- Next we look at concentration, for sums of bounded, and unbounded random variables, as long as the tails are well controlled.
- Now you want uniform laws, or uniform error bounds. Why? Say you are looking at convergence of a nonconvex algorithm. You want to understand the behavior of the convergence within some radius of some local/global optima. Here is where uniform error bounds come in very handy.
- In order to do uniform laws, one also needs a handle over the expectations of the supremum. This is why we looked at:
  - Finite class lemma, VC dimension, Sauer's lemma
  - Covering and packing numbers, Chaining, metric entropy.
  - We also saw that covering numbers can be helpful in bounding tails of suprema, not just expectations.
  - As for distributional convergence, we only looked at the Hajek projections, which helped us with U statistics.