STAT 7200

Introduction to Advanced Probability
Lecture 18

Taylor R. Brown

1. Convergence Theorems
   - Exchange Different Operators
   - Moment Generating Functions
   - Large Deviations Theory

2. Theory of Convergence II
   - Weak Convergence

MGFs: recap

- **Moment generating function** of random variable X:

$$M_X(s) = \mathrm{E}(e^{sX}), s \in \mathrm{R}.$$

- If $X \perp Y$, then $M_{X+Y}(s) = M_X(s) M_Y(s)$.
- Let X be random variable such that $M_X(s) < \infty$ for $0 < |s| < s_0$. Then $\mathrm{E}(|X^n|) < \infty$ for all n. And for $|s| < s_0$, we have:

$$M_X(s) = \sum_{k=0}^{\infty} \mathrm{E}(X^k) s^k / k! \text{ and } \mathrm{E}(X^r) = M_X^{(r)}(0)$$

.

Large Deviations Theory

- LLN:, for i.i.d. rvs $X_1, X_2, \ldots$, with finite mean μ and finite variance σ^2: for any $\varepsilon > 0$,

$$\mathrm{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0.$$

- However, this does not tell us any thing about the **speed** of convergence. It does not tell us how large n should be to guarantee the probability of a large deviation would be smaller than a certain threshold (say 5%).
- **Large deviations theory** studies the speed of convergence by estimating the probability of the large deviation as a function of sample size n.

Large Deviations Theory and Moment Generating functions

- Cherbychev's inequality gives us a weak result:

$$\mathsf{P}\left(\frac{X_1+\cdots+X_n}{n}\geq\mu+\varepsilon\right)\leq\mathsf{P}\left(\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|\geq\varepsilon\right)\leq\frac{\sigma^2}{n\varepsilon}$$

That is, the probability of large deviation decreases in the order of $O(1/n)$ (roughly proportional to $1/n$), which is rather slow. But this theorem is applicable under broad conditions.

Large Deviations Theory and Moment Generating functions

Theorem 1 (Large Deviation Theorem for LLN)

Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ, and $M_{X_i}(s) < \infty$ for $s \in (-a, b)$ where $a, b > 0$. Then

$$\mathrm{P}\left(\frac{X_1 + \cdots + X_n}{n} \geq \mu + \varepsilon\right) \leq \rho^n$$

where $\rho = \inf_{0<s<b}[e^{-s(\mu+\varepsilon)} M_{X_1}(s)] < 1$.

Large Deviation Theorem for LLN: Proof

- **Proof:** To estimate $P(\frac{X_1+\cdots+X_n}{n} \geq \mu+\varepsilon)$, let $Y_i = X_i - \mu - \varepsilon$, then for $-a < s < b$, we have $M_{Y_i}(s) = e^{-s(\mu+\varepsilon)} M_{X_i}(s) < \infty$. Also note that for $s > 0$, the function e^{sx} is an increasing and non-negative function, then by Markov's inequality:

$$
\begin{aligned}
& P(\frac{X_1+\cdots+X_n}{n} \geq \mu+\varepsilon) \\
=& P(\frac{Y_1+\cdots+Y_n}{n} \geq 0) = P(Y_1+\cdots+Y_n \geq 0) \\
=& P(e^{s(Y_1+\cdots+Y_n)} \geq 1) \leq M_{Y_1+\cdots Y_n}(s) = [e^{-s(\mu+\varepsilon)} M_{X_1}(s)]^n
\end{aligned}
$$

  for all $0 < s < b$. Thus, the probability of large deviation should be bounded by ρ^n where $\rho = \inf_{0<s<b}[e^{-s(\mu+\varepsilon)} M_{X_1}(s)]$.

- We still need to show that $\rho < 1$. Define $g(s) = e^{-s(\mu+\varepsilon)} M_{X_1}(s)$, we can verify that $g(0) = 1$ and $g'(0) = -\varepsilon < 0$. This suggests that, the function $g(s)$ would decrease as s increases around the $s = 0$. Then for some small $s > 0$, we must have $g(s) < 1$. As a result, $\rho < 1$.

Large Deviations: Example

- Let us consider the case when $X_1, X_2, \ldots$ are i.i.d. standard normal random variables. Let us take $\varepsilon = 1$. Without using the mgf, we have:

$$\mathrm{P}(\frac{X_1 + \cdots + X_n}{n} \geq 1) \leq \frac{1}{n} \leq \alpha$$

- By the large deviation theorem we just stated, and note that the mgf of $N(0,1)$ is $e^{s^2/2}$ for all s, we have:

$$\mathrm{P}\left(\frac{X_1 + \cdots + X_n}{n} \geq 1\right) \leq [\inf_{s>0} e^{-s} e^{s^2/2}]^n = e^{-\frac{n}{2}} \leq \alpha.$$

  This result suggests that, if we want to control the probability of large (> 1) deviation under α, then n must be at least $2\log(1/\alpha)$

- For $\alpha = 0.05$, the first estimation yields $n = 20$ while the second estimation yields $n = 6$. For $\alpha = 0.01$, the first estimation yields $n = 100$ while the second estimation yields $n = 10$. If $\alpha = 0.001$, the results are $n = 1000$ and $n = 14$ respectively.

Weak Convergence (chapter 10)

3.1: Definition: Weak Convergence

Given Borel probability measures $\mu_1, \mu_2, \dots,$ on R, we say that $\{\mu_n\}$ converges weakly to μ if $\int_{\mathsf{R}} f d\mu_n \to \int_{\mathsf{R}} f d\mu$ for *all bounded continuous functions* $f : \mathsf{R} \to \mathsf{R}$. (Also known as convergence in distribution).

Weak Convergence (chapter 10)

- **Example:** Let $\mu_n \sim N(0, \frac{1}{n})$, then $\{\mu_n\}$ converges weakly to δ_0
- **Example:** Let $\mu_n \sim N(0, \frac{n-1}{n})$, then $\{\mu_n\}$ converges weakly to $N(0,1)$
- **Note:** Unlike the convergence almost surely and convergence in probability, which are about the convergence of random variables, the weak convergence solely focuses on the convergence of probability measures. Thus, the weak convergence of measure $\{\mu_n\}$ usually does not guarantee the convergence of random variables $\{X_n\}$ even if $\mathcal{L}(X_n) \sim \mu_n$.
- For instance, for random variable $Z \sim N(0,1)$, define $Z_k = Z$ when k is odd, and $Z_k = -Z$ when k is even. Then $\{Z_n\}$ does not converge almost surely or in probability. However, since $\mathcal{L}(Z_n) \sim N(0,1)$, the distributions of $\{Z_n\}$ would converge weakly to $N(0,1)$.

Equivalent Definitions of Weak Convergence

Theorem 2 (Equivalent Definitions of Weakly Convergence)

The following statements are all equivalent definitions:
(1) $\{\mu_n\}$ converges weakly to μ. (Original definition)

(2) $\mu_n(A) \to \mu(A)$ for all measurable set A such that $\mu(\partial A) = 0$. (∂A is defined as the boundary of set A)

(3) $\mu_n((-\infty, x]) \to \mu((-\infty, x])$ for all $x \in \mathrm{R}$ such that $\mu(\{x\}) = 0$. That is, the convergence of CDFs. (Note, $\{x\}$ is the boundary of set $(-\infty, x]$.)

(4) (Skorohod's Theorem) there are rvs $Y, Y_1, Y_2, \dots$ defined on the same probability triple, with $\mathcal{L}(Y) = \mu$ and $\mathcal{L}(Y_n) = \mu_n$ such that $Y_n \to Y$ with probability 1 (This theorem connects the strongest type of convergence: convergence almost surely, with weak convergence.)

(5) $\int_{\mathrm{R}} f d\mu_n \to \int_{\mathrm{R}} f d\mu$ for all bounded Borel-measurable functions $f : \mathrm{R} \to \mathrm{R}$. such that $\mu(D_f) = 0$, where D_f is the set of discontinuous points of f. (The continuous condition of definition 1) is relaxed.)

Proof: Some Immediate Results

- (1) $\{\mu_n\}$ converges weakly to μ.

  (2) $\mu_n(A) \to \mu(A)$ for all measurable set A such that $\mu(\partial A) = 0$.

  (5) $\int_R f d\mu_n \to \int_R f d\mu$ for all bounded Borel-measurable functions $f : R \to R$. such that $\mu(D_f) = 0$, where D_f is the set of discontinuous points of f.

- (5) $\Rightarrow$ (1): Immediate result, since the set of discontinuous points of a continuous function is the empty set.
- (5) $\Rightarrow$ (2): Let $f = 1_A$, then the set of discontinuous points of f is the boundary of A, so $D_{1_A} = \delta A$, $\mu(D_{1_A}) = \mu(\delta A) = 0$, $\mu_n(A) = \int_R f d\mu_n \to \int_R f d\mu = \mu(A)$.

Proof: Some Immediate Results

- (2) $\mu_n(A) \to \mu(A)$ for all measurable set A such that $\mu(\partial A) = 0$.

  (3) $\mu_n((-\infty, x]) \to \mu((-\infty, x])$ for all $x \in \mathrm{R}$ such that $\mu(\{x\}) = 0$.

- (2) $\Rightarrow$ (3): Immediate result, since $\partial(-\infty, x] = \{x\}$.
- **Interior, boundary, closure and exterior:**
  Given any set $A \subseteq \mathrm{R}$, the interior is defined as

$$\{x \in A : \exists \varepsilon > 0, (x - \varepsilon, x + \varepsilon) \subseteq A\},$$

  the boundary is defined as

$$\partial A = \{x \in \mathrm{R} : \forall \varepsilon > 0, A \cap (x - \varepsilon, x + \varepsilon) \neq \emptyset, A^c \cap (x - \varepsilon, x + \varepsilon) \neq \emptyset\},$$

  the closure is defined as the union of A and its boundary, and exterior is defined as the interior of the complement of A. Interior, boundary and exterior forms a partition of R.

- For instance, the interior of $(a, b]$ is (a, b), the boundary is $\{a, b\}$, the closure is $[a, b]$ and the exterior is $(-\infty, a) \cup (b, \infty)$.