Lecture 2: Expectation

Zhentao Shi

August 10, 2017

Expected Value

Integration

Integration is one of the most fundamental operations in mathematical analysis. We have studied Riemann's integral in the undergraduate calculus. Riemann's integral is intuitive, but it encounters difficulty when it handles infinite value of the integrand or the region.

Lebesgue integral is a more general way to define integration. It is constructed by the following steps. X is called a *simple function* on a measurable space $(\Omega, \mathcal{F})$ if $X = \sum_i a_i \cdot 1\left\{A_i\right\}$ is a finite sum, where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$.

1. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space. The integral of the simple function X with respect to μ is
$$\int X \mathrm{d}\mu = \sum_i a_i \mu\left(A_i\right).$$
2. Let X be a non-negative measurable function. The integral of X with respect to μ is
$$\int X \mathrm{d}\mu = \sup\left\{\int Y \mathrm{d}\mu : 0 \leq Y \leq X,\ Y \text{ is simple}\right\}.$$
3. Let X be a measurable function. Define $X^+ = \max\left\{X, 0\right\}$ and $X^- = -\min\left\{X, 0\right\}$. Both X^+ and X^- are non-negative functions. The integral of X with respect to μ is
$$\int X \mathrm{d}\mu = \int X^+ \mathrm{d}\mu - \int X^- \mathrm{d}\mu.$$

Step 1 defines the integral of a simple function. Step 2 defines the integral of a non-negative function as the approximation of steps functions from below. Step 3 defines the integral of a general function as the difference of the integral of two non-negative parts.

If the measure μ is a probability measure P, then the integral $\int X \mathrm{d}P$ is called the *expected value,* or *expectation,* of X. We often use the notation $E[X]$, instead of $\int X \mathrm{d}P$, for convenience. Expectation is the average of a random variable.

If we know the probability mass function of a discrete random variable, its expectation is calculated as $E[X] = \sum_x xP(X = x)$, which is the integral of a simple function. If a continuous random variable has a PDF, its expectation is computed as $E[X] = \int xf(x) \mathrm{d}x$.

Here are some properties of the expectation.

- The probability of an event A is the expectation of an indicator function. $E[1\{A\}] = 1 \times P(A) + 0 \times P(A^c) = P(A)$.

- $E[X^r]$ is call the r-moment of X. The mean of a random variable is the first moment $\mu = E[X]$, and the second *centered* moment is called the variance $\mathrm{var}[X] = E\left[(X - \mu)^2\right]$. The third centered moment $E\left[(X - \mu)^3\right]$, called skewness, is a measurement of the symmetry of a random variable, and the fourth centered moment $E\left[(X - \mu)^4\right]$, called kurtosis, is a measurement of the tail thickness.

- We call $E\left[(X - \mu)^3\right]/\sigma^3$ the *skewness coefficient*, and $E\left[(X - \mu)^4\right]/\sigma^4 - 3$ *degree of excess*. A normal distribution's skewness and degree of excess are both zero.

  - **Application**: The formula that killed Wall Street

- $E[\cdot]$ is a linear operation. If $\phi(\cdot)$ is a linear function, then $E[\phi(X)] = \phi(E[X])$.

- *Jensen's inequality* is an important fact. A function $\varphi(\cdot)$ is convex if $\varphi(ax_1 + (1-a)x_2) \leq a\varphi(x_1) + (1-a)\varphi(x_2)$ for all x_1, x_2 in the domain and $a \in [0,1]$. For instance, x^2 is a convex function. Jensen's inequality says that if $\varphi(\cdot)$ is a convex function, then $\varphi(E[X]) \leq E[\varphi(X)]$.

- **Application**: The *Kullback-Leibler distance* is defined as $d(p,q) = \int \log(p/q)\,\mathrm{d}P = E_P[\log(p/q)]$ for two probability measures p and q. The Kullback-Leibler distance $d(p,q) \geq 0$ and the inequality holds if and only if $p = q$.

- *Markov inequality* is another simple but important fact. If $E[|X|^r]$ exists, then $P(|X| > \epsilon) \leq E[|X|^r]/\epsilon^r$ for all $r \geq 1$. *Chebyshev inequality* $P(|X| > \epsilon) \leq E[X^2]/\epsilon^2$ is a special case of the Markov inequality when $r = 2$.

Multivariate Random Variable

A bivariate random variable is a measurable function $X : \Omega \mapsto \mathbb{R}^2$, and more generally a multivariate random variable is a measurable function $X : \Omega \mapsto \mathbb{R}^n$. We can define the *joint CDF* as $F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$. Joint PDF is defined similarly.

It is sufficient to introduce the joint distribution, conditional distribution and marginal distribution in the simple bivariate case, and these definitions can be extended to multivariate distributions. Suppose a bivariate random variable (X, Y) has a joint density $f(\cdot, \cdot)$. The *conditional density* can be roughly written as $f(y|x) = f(x,y)/f(x)$ if we do not formally deal with the case $f(x) = 0$. The *marginal density* $f(y) = \int f(x,y)\,dx$ integrates out the coordinate that is not interested.

Independence

In a probability space $(\Omega, \mathcal{F}, P)$, for two events $A_1, A_2 \in \mathcal{F}$ the *conditional probability* is

$$P(A_1|A_2) = \frac{P(A_1 A_2)}{P(A_2)}$$

if $P(A_2) \neq 0$. If $P(A_2) = 0$, the conditional probability can still be valid in some cases, but we need to introduce the *dominance* between two measures, which I choose not to do at this time. In the definition of conditional probability, A_2 plays the role of the outcome space so that $P(A_1A_2)$ is standardized by the total mass $P(A_2)$.

Since A_1 and A_2 are symmetric, we also have $P(A_1A_2) = P(A_2|A_1)P(A_1)$. It implies

$$P(A_1|A_2) = \frac{P(A_2|A_1)\, P(A_1)}{P(A_2)}$$

This formula is the well-known *Bayes' Theorem*. It is particularly important in decision theory.

Example: A_1 is the event "a student can survive CUHK's MSc program", and A_2 is his or her application profile.

We say two events A_1 and A_2 are *independent* if $P(A_1A_2) = P(A_1)P(A_2)$. If $P(A_2) \neq 0$, it is equivalent to $P(A_1|A_2) = P(A_1)$. In words, knowing A_2 does not change the probability of A_1.

Regarding the independence of two random variables, X and Y are *independent* if $P(X \in B_1, Y \in B_2) = P(X \in B_1)\, P(Y \in B_2)$ for any two Borel sets B_1 and B_2.

If X and Y are independent, $E[XY] = E[X]E[Y]$.

Application: (Chebyshev law of large numbers) If $X_1, X_2, \ldots, X_n$ are independent, and they have the same mean 0 and variance $\sigma^2 < \infty$. Let $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then the probability $P(|Z_n| > \epsilon) \to 0$ as $n \to \infty$,

Law of Iterated Expectations

Given a probability space $(\Omega, \mathcal{F}, P)$, a sub σ-algebra $\mathcal{G} \subset \mathcal{F}$ and a $\mathcal{F}$-measurable function X with $E|X| < \infty$, the *conditional expectation* $E[X|\mathcal{G}]$ is defined as a $\mathcal{G}$-measurable function such that $\int_A X dP = \int_A E[X|\mathcal{G}]\, dP$ for all $A \in \mathcal{G}$. *Law of iterated expectation* is a trivial fact if we take $A = \Omega$.

In the bivariate case, if the conditional density exists, the conditional expectation can be computed as $E[Y|X] = \int yf(y|X)\,dy$. The law of iterated expectation implies $E[E[Y|X]] = E[Y]$.

Below are some properties of conditional expectations

1. $E[E[Y|X_1, X_2]|X_1] = E[Y|X_1]$;
2. $E[E[Y|X_1]|X_1, X_2] = E[Y|X_1]$;
3. $E[h(X)Y|X] = h(X)E[Y|X]$.

Application: Regression is a technique that decomposes a random variable Y into two parts, a conditional mean and a residual. Write $Y = E[Y|X] + \epsilon$, where $\epsilon = Y - E[Y|X]$. Show that $E[\epsilon] = 0$ and $E[\epsilon E[Y|X]] = 0$.