5. Expectation

Integration. (Brief Review)

First recall that the *indicator function* of a measurable set E is defined as

$$1_E(x) = \begin{cases} 1, & x \in E \\ 0, & x \notin E \end{cases},$$

and a *simple function* $\phi = \sum_{i=1}^n a_i 1_{E_i}$ is a linear combination of indicator functions (where we may assume that the coefficients are distinct).

A fundamental observation is that we can approximate a measurable function with simple functions by partitioning the codomain.

Theorem 5.1. *If $(S, \mathcal{G})$ is a measurable space and $f : S \to [0, \infty]$ is measurable, then there is a sequence $\{\phi_n\}_{n=1}^\infty$ of simple functions with $0 \le \phi_1 \le \phi_2 \le ... \le f$ such that $\phi_n \to f$ pointwise, and the convergence is uniform on any set on which f is bounded.*

Proof. For $n = 1, 2, ...$ and $k = 0, 1, ..., 4^n - 1$, define

$$E_n^k = f^{-1}\left(\left(\frac{k}{2^n}, \frac{k+1}{2^n}\right]\right) \text{ and } F_n = f^{-1}\left((2^n, \infty]\right),$$

and set

$$\phi_n = \sum_{k=0}^{4^n-1} \frac{k}{2^n} 1_{E_n^k} + 2^n 1_{F_n}.$$ □

Now let $(S, \mathcal{G}, \mu)$ be a measure space. We construct the integral as follows:

(i) For any $E \in \mathcal{G}$,

$$\int 1_E d\mu = \mu(E).$$

(ii) For any simple function $\phi = \sum_{i=1}^n a_i 1_{E_i}$,

$$\int \phi \, d\mu = \sum_{i=1}^n a_i \int 1_{E_i} d\mu = \sum_{i=1}^n a_i \mu(E_i)$$

with the convention that $0 \cdot \infty = 0$.

(iii) For any measurable function $f : S \to [0, \infty]$,

$$\int f d\mu = \sup\left\{\int \phi \, d\mu : 0 \le \phi \le f, \ \phi \text{ is simple}\right\}.$$

(This is equal to $\lim_{n\to\infty} \int \phi_n d\mu$ with ϕ_n as in the proof of Theorem 5.1 by the MCT.)

(iv) For any measurable $f : S \to \overline{\mathbb{R}}$ with $\int |f| \, d\mu < \infty$ (called an *integrable function*),

$$\int f d\mu = \int (f \vee 0) d\mu - \int (-f \vee 0) d\mu.$$

For f integrable, $A \in \mathcal{G}$, we define the integral of f over A as $\int_A f d\mu = \int f 1_A d\mu$.

When we wish to emphasize dependence on the argument, we write $\int f d\mu = \int f(x) d\mu(x)$, or sometimes $\int f(x) \mu(dx)$.

Proposition 5.1. *For any $a, b \in \mathbb{R}$ and any integrable functions f, g, $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$. If $f \le g$ a.e., then $\int f d\mu \le \int g d\mu$.*

Definition.

If X is a random variable on $(\Omega, \mathcal{F}, P)$ with $X \geq 0$ a.s., then we define its expectation as $E[X] = \int X dP$, which always makes sense, but may be $+\infty$.

If X is an arbitrary random variable, then we can write $X = X^+ - X^-$ where $X^+ = \max\{0, X\}$ and $X^- = \max\{0, -X\}$ are nonnegative random variables.

If at least one of $E[X^+], E[X^-]$ is finite, then we define $E[X] = E[X^+] - E[X^-]$.

Note that $E[X]$ may be defined even if X isn't an integrable function.

A trivial but extremely useful observation is that $P(A) = E[1_A]$ for any event $A \in \mathcal{F}$.

Inequalities.

Recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be *convex* if for every $x, y \in \mathbb{R}$, $\lambda \in [0, 1]$, we have

$$\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

That is, given any two points $x, y \in \mathbb{R}$, the line from $(x, \varphi(x))$ to $(y, \varphi(y))$ lies above the graph of φ.

Lemma 5.1. *If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then*

$$\frac{\varphi(y) - \varphi(x)}{y - x} \leq \frac{\varphi(z) - \varphi(x)}{z - x} \leq \frac{\varphi(z) - \varphi(y)}{z - y}$$

for every $x < y < z$.

Proof. (Homework)

Writing $\lambda = \frac{y-x}{z-x} \in (0, 1)$, we have $y = \lambda z + (1-\lambda)x$, so it follows from convexity that
$\varphi(y) \leq \lambda\varphi(z) + (1-\lambda)\varphi(x)$, and thus

$$\varphi(y) - \varphi(x) \leq \lambda\left(\varphi(z) - \varphi(x)\right) = \frac{y-x}{z-x}\left(\varphi(z) - \varphi(x)\right).$$

Dividing by $y - x > 0$ gives the first inequality.

Similarly, setting $\mu = \frac{z-y}{z-x} = 1 - \lambda \in (0, 1)$, we have $y = \mu x + (1-\mu)z$, so $\varphi(y) \leq \mu\varphi(x) + (1-\mu)\varphi(z)$, and thus

$$\varphi(y) - \varphi(z) \leq \mu\left(\varphi(x) - \varphi(z)\right) = \frac{z-y}{z-x}\left(\varphi(x) - \varphi(z)\right),$$

hence

$$\frac{\varphi(z) - \varphi(y)}{z - y} \geq \frac{\varphi(z) - \varphi(x)}{z - x}.$$

□

Lemma 5.2 (Supporting Hyperplane Theorem in $\mathbb{R}^2$)**.** *If φ is a convex function, then for any $c \in \mathbb{R}$, there is a linear function $l(x)$ which satisfies $l(c) = \varphi(c)$ and $l(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$.*

Proof. (Homework)

For any $h > 0$, taking $x = c - h$, $y = c$, $z = c + h$ in Lemma 5.1, it follows from the outer inequality that that

$$\frac{\varphi(c) - \varphi(c-h)}{h} \leq \frac{\varphi(c+h) - \varphi(c)}{h}.$$

Also, for any $0 < h_1 < h_2$, we have $c - h_2 < c - h_1 < c$, so the second inequality in Lemma 5.1 shows that $\frac{\varphi(c)-\varphi(c-h_2)}{h_2} \leq \frac{\varphi(c)-\varphi(c-h_1)}{h_1}$.

Similarly, since $c < c+h_1 < c+h_2$, the first inequality in Lemma 5.1 shows that $\frac{\varphi(c+h_2)-\varphi(c)}{h_2} \geq \frac{\varphi(c+h_1)-\varphi(c)}{h_1}$. Consequently, the one-sided derivatives exist and satisfy

$$\varphi_l'(c) := \lim_{h\to 0^+} \frac{\varphi(c)-\varphi(c-h)}{h} \leq \lim_{h\to 0^+} \frac{\varphi(c+h)-\varphi(c)}{h} := \varphi_r'(c).$$

Now let $a \in [\varphi_l'(c), \varphi_r'(c)]$ and define the linear function $l(x) = a(x-c)+\varphi(c)$. Clearly, $l(c) = \varphi(c)$. To see that $l(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$, note that if $x < c$, then $x = c-k$ for some $k > 0$, so

$$l(x) - \varphi(x) = a(x-c) + \varphi(c) - \varphi(c-k) = -k\left(a - \frac{\varphi(c)-\varphi(c-k)}{k}\right) \leq 0$$

since $\frac{\varphi(c)-\varphi(c-k)}{k} \leq \varphi_l'(c) \leq a$ by monotonicity. The $x > c$ case is similar. □

Theorem 5.2 (Jensen). *If φ is a convex function and X is a random variable, then*

$$\varphi\left(E[X]\right) \leq E\left[\varphi(X)\right]$$

whenever the expectations exist.

Proof. Lemma 5.2 gives the existence of a function $l(x) = ax+b$ which satisfies $l\left(E[X]\right) = \varphi\left(E[X]\right)$ and $l(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$.

By monotonicity and linearity, we have

$$\begin{aligned} E[\varphi(X)] &= \int \varphi(X)dP \geq \int l(X)dP = \int (aX+b)dP \\ &= a\int XdP + b = aE[X]+b = l\left(E[X]\right) = \varphi\left(E[X]\right). \end{aligned}$$ □

The triangle inequality $E|X| \geq |E[X]|$ is an important special case.

I remember the direction in Jensen's inequality by $E\left[X^2\right] - E[X]^2 = \mathrm{Var}(X) \geq 0$.

A function is called *strictly convex* if the defining inequality is strict. For such functions, modifying the preceding arguments where necessary shows that Jensen's inequality is strict unless X is a.s. constant.

To state the next inequality, we define the L^p norm of a random variable by $\|X\|_p = E\left[|X|^p\right]^{\frac{1}{p}}$ for $p \in [1,\infty)$ and $\|X\|_\infty = \inf\{M : P(|X| > M) = 0\}$.

We define $L^p = L^p(\Omega, \mathcal{F}, P) = \left\{X : \|X\|_p < \infty\right\}$ (where random variables X and Y define the same element of L^p if they are equal almost surely), and one can prove that L^p is a Banach space for $p \geq 1$.

Theorem 5.3 (Hölder). *If $p, q \in [1,\infty]$ with $\frac{1}{p}+\frac{1}{q} = 1$ (where $\frac{1}{\infty} = 0$), then*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

Proof.

We first note that the result holds trivially if the right-hand side is infinity, and if $\|X\|_p = 0$ or $\|Y\|_q = 0$, then $|XY| = 0$ a.s. Accordingly, we may assume that $0 < \|X\|_p, \|Y\|_q < \infty$. In fact, since constants factor out of L^p-norms, it suffices to establish the result when $\|X\|_p = \|Y\|_q = 1$.

Also, the case $p = \infty, q = 1$ (and symmetrically) is immediate since $|X| \leq \|X\|_\infty$ a.s., thus

$$E|XY| \leq E\left[\|X\|_\infty |Y|\right] = \|X\|_\infty E|Y| = \|X\|_\infty \|Y\|_1.$$

Accordingly, we will assume henceforth that $p, q \in (1, \infty)$.

Now fix $y \geq 0$, and define the function $\varphi : [0, \infty) \to \mathbb{R}$ by $\varphi(x) = \frac{x^p}{p} + \frac{y^q}{q} - xy$.

Since $\varphi'(x) = x^{p-1} - y$ and $\varphi''(x) = (p-1)x^{p-2} > 0$ for $x > 0$, φ attains its minimum at $x_0 = y^{\frac{1}{p-1}}$.

Thus, as the conjugacy of p and q implies that $\frac{1}{p-1} + 1 = \frac{p}{p-1} = \left(1 - \frac{1}{p}\right)^{-1} = q$, we have that

$$\varphi(x) \geq \varphi(x_0) = \frac{x_0^p}{p} + \frac{y^q}{q} - xy = \frac{y^{\frac{p}{p-1}}}{p} + \frac{y^q}{q} - y^{\frac{1}{p-1}+1} = y^q \left(\frac{1}{p} + \frac{1}{q}\right) - y^q = 0$$

for all $x \geq 0$. It follows that $\frac{x^p}{p} + \frac{y^q}{q} \geq xy$ for every $x, y \geq 0$.

In particular, taking $x = |X|$, $y = |Y|$, and integrating, we have

$$\begin{aligned} E|XY| &= \int |X||Y| \, dP \leq \frac{1}{p} \int |X|^p \, dP + \frac{1}{q} \int |Y|^q \, dP \\ &= \frac{\|X\|_p^p}{p} + \frac{\|Y\|_q^q}{q} = \frac{1}{p} + \frac{1}{q} = 1 = \|X\|_p \|Y\|_q . \end{aligned}$$

□

Some useful corollaries of Hölder's inequality are:

Corollary 5.1 (Cauchy-Schwarz)**.** $E|XY| \leq \sqrt{E[X^2] E[Y^2]}$.

Alternate Proof. For all $t \in \mathbb{R}$,

$$0 \leq E\left[(|X| + t|Y|)^2\right] = E\left[X^2\right] + 2tE|XY| + t^2 E\left[Y^2\right] = q(t),$$

so $q(t)$ has at most one real root and thus a nonpositive discriminant

$$(2E|XY|)^2 - 4E\left[X^2\right] E\left[Y^2\right] \leq 0.$$

□

Corollary 5.2. *For any random variable X and any $1 \leq r < s \leq \infty$, $\|X\|_r \leq \|X\|_s$. Therefore, we have the inclusion $L^s \subseteq L^r$.*

Proof. For $s = \infty$, we have $|X|^r \leq \|X\|_\infty^r$ a.s., hence

$$\|X\|_r^r = \int |X|^r \, dP \leq \int \|X\|_\infty^r \, dP = \|X\|_\infty^r .$$

For $s < \infty$, apply Holder's inequality to X^r and 1 with $p = \frac{s}{r}$, $q = \frac{s}{s-r}$ to get

$$\|X\|_r^r = E[|X|^r] \leq \|X^r\|_p \|1\|_q = \left(\int |X^r|^{\frac{s}{r}} \, dP\right)^{\frac{r}{s}} = \|X\|_s^r .$$

□

Note that for Corollary 5.2, it is important that our measure was finite.

Of course, we could also prove the inclusion by breaking up the integral according to whether $|X|$ is greater or less than 1, though we would not obtain the inequality in that case.

The proof of our last big inequality should be familiar from measure theory (convergence in L^1 implies convergence in measure).

Theorem 5.4 (Chebychev)**.** *For any nonnegative random variable X and any $a > 0$,*

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

Proof. Let $A = \{\omega : X(\omega) \geq a\}$. Then

$$aP(X \geq a) = a \int 1_A dP \leq \int X 1_A dP \leq \int X dP = E[X]. \qquad \square$$

Corollary 5.3. *For any $(S, \mathcal{G})$-valued random variable X and any measurable function $\varphi : S \to [0, \infty)$,*

$$P(\varphi(X) \geq a) \leq \frac{E[\varphi(X)]}{a}.$$

Some important cases of Corollary 5.3 for real-valued X are

- $\varphi(x) = |x|$: to control the probability that an integrable random variable is large.
- $\varphi(x) = (x - E[X])^2$: to control the probability that a random variable with finite variance is far from its mean.
- $\varphi(x) = e^{tx}$: to establish exponential decay for random variables with moment generating functions (*concentration inequalities*).

Limit Theorems.

We now briefly recall the three main results for interchanging limits and integration. The proofs can be found in any book on measure theory.

Theorem 5.5 (Monotone Convergence Theorem)**.** *If $0 \leq X_n \nearrow X$, then $E[X_n] \nearrow E[X]$.*

Theorem 5.6 (Fatou's Lemma)**.** *If $X_n \geq 0$, then $\liminf_{n\to\infty} E[X_n] \geq E\left[\liminf_{n\to\infty} X_n\right]$.*

Note that if $X_n \geq M$, then $X_n - M \geq 0$, so since constants behave nicely with respect to limits and expectation, "nonnegative" can be relaxed to "bounded below" in the statement of Theorems 5.5 and 5.6.

Also, since $X_n \nearrow X$ if and only if $(-X_n) \searrow (-X)$ and $\liminf_{n\to\infty} X_n = -\limsup_{n\to\infty}(-X_n)$, one has immediate corollaries for lim sups and for monotone decreasing sequences (provided that they are bounded above).

Theorem 5.7 (Dominated Convergence Theorem)**.** *If $X_n \to X$ and there exists some $Y \geq 0$ with $E[Y] < \infty$ and $|X_n| \leq Y$ for all n, then $E[X_n] \to E[X]$.*

When Y is a constant, Theorem 5.7 is known as the bounded convergence theorem. In that case, it is important that we're working on a finite measure space.

In each of these theorems, the assumptions need only hold almost surely since one can modify random variables on null sets without affecting their expectations.

Change of Variables.

Though integration over arbitrary measure spaces in nice in theory, in order to actually compute expectations, we will typically need to work in more familiar spaces like $\mathbb{R}^d$.

The following change of variables theorem allows us to compute expectations by integrating functions of a random variable against its distribution.

Theorem 5.8. *Let X be a random variable taking values in the measurable space $(S, \mathcal{G})$, and let $\mu = P \circ X^{-1}$ be the pushforward measure on $(S, \mathcal{G})$.*
If f is a measurable function from $(S, \mathcal{G})$ to $(\mathbb{R}, \mathcal{B})$ such that $f \geq 0$ or $E|f(X)| < \infty$, then

$$E[f(X)] = \int_S f(s) d\mu(s).$$

Proof. We will proceed by verifying the result in increasingly general cases paralleling the construction of the integral.

To begin with, let $B \in \mathcal{G}$ and $f = 1_B$. Then

$$E[f(X)] = E[1_B(X)] = P(X \in B) = \mu(B) = \int_S 1_B(s) d\mu(s) = \int_S f(s) d\mu(s).$$

Now suppose that $f = \sum_{i=1}^n a_i 1_{B_i}$ is a simple function. Then by linearity and the previous case,

$$E[f(X)] = \sum_{i=1}^n a_i E[1_{B_i}(X)] = \sum_{i=1}^n a_i \int_S 1_{B_i}(s) d\mu(s) = \int_S f(s) d\mu(s).$$

If $f \geq 0$, then Theorem 5.1 gives a sequence of simple functions $\phi_n \nearrow f$, so the previous case and two applications of the MCT give

$$E[f(X)] = \lim_{n\to\infty} E[\phi_n(X)] = \lim_{n\to\infty} \int_S \phi_n(s) d\mu(s) = \int_S f(s) d\mu(s).$$

Finally, suppose that $E|f(X)| < \infty$, and set $f^+(x) = \max\{f(x), 0\}$, $f^-(x) = \max\{-f(x), 0\}$. Then $f^+, f^- \geq 0$, $f = f^+ - f^-$, and $E[f(X)^+], E[f(X)^-] \leq E|f(X)| < \infty$, so it follows from the previous result and linearity that

$$E[f(X)] = E[f^+(X)] - E[f^-(X)] = \int_S f^+(s) d\mu(s) - \int_S f^-(s) d\mu(s) = \int_S f(s) d\mu(s). \qquad \square$$

In light of Theorem 5.8, if X is an integrable random variable on $(\Omega, \mathcal{F}, P)$ with distribution μ, then

$$E[X] = \int X dP = \int_{\mathbb{R}} x d\mu(x).$$

If X has density $f = \frac{d\mu}{dm}$, then for any measurable $g : \mathbb{R} \to \mathbb{R}$ with $g \geq 0$ a.s. or $\int |g| \, d\mu < \infty$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

If X is a random variable, then for any $k \in \mathbb{N}$, we say that $E[X^k]$ is the *kth moment* of X.

The first moment $E[X]$ is called the *mean* and is usually denoted $E[X] = \mu$.
The mean is a measure of the center of the distribution of X.

If X has finite second moment $E[X^2] < \infty$, then we define the *variance* (or second central moment) of X as $\mathrm{Var}(X) = E[(X - \mu)^2]$.
The variance provides a measure of the dispersion of the distribution of X and is is usually denoted $\mathrm{Var}(X) = \sigma^2$.

By linearity, we have the useful identity

$$\mathrm{Var}(X) = E[(X - \mu)^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - E[X]^2.$$