STAT 512: Statistical Inference **Autumn 2022**

Quiz Session 1: Review of Prerequisites

Yikun Zhang *September 28, 2022*

Quiz Session:

- **Time:** Every Wednesday from 12:30pm to 1:20pm (there will be 11 quiz sessions in total).
- **Location:** Room 145 of Health Sciences Education Building.
- **Zoom Option:** `https://washington.zoom.us/j/98450014077?pwd=V1RIUlRCSTliK3R6NnRNZy80T1NxQT09` (password: 512quiz).
- **Attendance:** Recommended but not required.

Yikun's Office Hour:

- **Time:** Every Tuesday from 9:30am to 10:30am (with possible extension to 11am).
- **Location:** Padelford Hall B-226.
- **Virtual Option:** `https://washington.zoom.us/j/98194898781?pwd=bFpRSG4yclJETFNyNTcrR2RTcjhHZz09` (password: stat512).
- **Email:** yikun@uw.edu.

Zhen's Office Hour:

- **Time:** Every Wednesday from 4pm to 5pm.
- **Virtual Only:** `https://washington.zoom.us/j/99384780380`.
- **Email:** zhenm@uw.edu

Main Scheme of Quiz Sessions:

- Review a couple of key concepts in the lecture.
- Go through some homework problems that have been due in the previous weeks if necessary.
- Discuss other exercises that help strengthen the understanding of lecture materials or provide insights into other related study fields.

The first quiz session is intended to help assess your mathematical preparation for STAT 512 and review selected background knowledge.

Diagnostic Exercises: (check them by yourself and don't hand in!)

1. Evaluate $\lim\limits_{n\to\infty} p^{\frac{1}{n}}$ for any $p > 0$.
2. Evaluate $\sum\limits_{k=0}^{\infty} kx^k$ for $|x| < 1$.

3. Evaluate $\int_D \frac{1}{\pi}\, dxdy$ with $D = \left\{(x, y) \in \mathbb{R}^2 : \frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1\right\}$ being an ellipse, where $a, b > 0$.

4. Let X be a random variable uniformly distributed on the interval (a, b). Find $\mathbb{E}(X)$ and $\mathbb{P}[X \leq \mathbb{E}(X)]$.

5. Compute the determinant, eigenvalues, and eigenvectors of the matrix $\begin{bmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$.

1 Multivariate Calculus (or Mathematical Analysis)

Metric Space: A set $\mathbb{X}$, whose elements we shall call *points*, is said to be a *metric space* if with any two points p and q of $\mathbb{X}$ there is associated a real number $d(p, q)$, called the *distance* from p to q, such that

(a) $d(p, q) > 0$ if $p \neq q$ and $d(p, p) = 0$ for any $p \in \mathbb{X}$;

(b) $d(p, q) = d(q, p)$;

(c) $d(p, q) \leq d(p, r) + d(r, q)$ for any $r \in \mathbb{X}$.

Any function with these three properties is called a *distance function*, or *metric.*

Example 1 (Euclidean Spaces)**.** *In STAT 512, the major metric space of concern will be the Euclidean spaces $\mathbb{R}^n$ for some integer n, whose distance is defined by*

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2 \qquad \text{for any } \boldsymbol{x} = (x_1, ..., x_n)^T, \boldsymbol{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n, \tag{1}$$

in which $||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^n x_i^2}$ is the L_2 norm *in $\mathbb{R}^n$ and the superscript T stands for the (vector/matrix) transpose. Notice that L_2 norm can be determined by the* inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^n x_i y_i \tag{2}$$

in $\mathbb{R}^n$. Other possible norms in $\mathbb{R}^n$ include

- L_p norm*: $||\boldsymbol{x}||_p = (\sum_{i=1}^n x_i^p)^{\frac{1}{p}}$;*
- Infinity norm*: $||\boldsymbol{x}||_\infty = \max_i |x_i|$.*

Convergent Sequence: A sequence $\{p_n\}$ in a metric space $\boldsymbol{X}$ is said to *converge* if there is a point $p \in \mathbb{X}$ with the following property: For every $\epsilon > 0$ there is an integer N such that $n \geq N$ implies that $d(p_n, p) < \epsilon$. We write $p_n \to p$ or $\lim_{n\to\infty} p_n = p$ and call p the limit of $\{p_n\}$.

Subsequence: Given a sequence $\{p_n\}$, consider an increasing sequence $\{n_k\}$ of positive integers. Then the sequence $\{p_{n_k}\}$ is called a *subsequence* of $\{p_n\}$. If $\{p_{n_k}\}$ converges, its limit is called a *sequential limit* of $\{p_n\}$.

Cauchy Sequence: A sequence $\{p_n\}$ in a metric space $\mathbb{X}$ is said to be *Cauchy sequence* if for every $\epsilon > 0$ there is an integer N such that $d(p, p_n) < \epsilon$ if $n \geq N$ and $m \geq N$.

Proposition 1. *Let $\{p_n\}$ be a sequence in a metric space $\mathbb{X}$.*

(a) Whenever $\{p_n\}$ converges, its limit is unique.

(b) *If $\{p_n\}$ converges, then $\{p_n\}$ is bounded,* i.e., *there exists a real number $M > 0$ and a point $q \in \mathbb{X}$ such that $d(p_n, q) < M$ for all n.*

(c) *Every bounded sequence in $\mathbb{R}^n$ contains a convergent subsequence.*

(d) *In any metric space $\mathbb{X}$, every convergent sequence is a Cauchy sequence.*

(e) *In $\mathbb{R}^n$, every Cauchy sequence converges.*

Proof. See Theorem 3.2, 3.6, 3.11 in Rudin [1976]. □

Lower and Upper Limits: Let $\{s_n\}$ be a sequence of real numbers. The *lower limit* or *limit inferior* of $\{s_n\}$ is defined by

$$\liminf_{n\to\infty} s_n := \lim_{n\to\infty}\left(\inf_{m\geq n} s_m\right) \equiv \sup\left\{\inf\left\{s_m : m \geq n\right\} : n \geq 0\right\}. \tag{3}$$

Similarly, the *upper limit* or *limit superior* of $\{s_n\}$ is defined by

$$\limsup_{n\to\infty} s_n := \lim_{n\to\infty}\left(\sup_{m\geq n} s_m\right) \equiv \inf\left\{\sup\{s_m : m \geq n\} : n \geq 0\right\}. \tag{4}$$

Equivalently, the lower and upper limits can be defined through the subsequential limits of $\{s_n\}$ as

$$\liminf_{n\to\infty} s_n = \inf E \quad \text{and} \quad \limsup_{n\to\infty} s_n = \sup E,$$

where $E \subseteq [-\infty, \infty]$ is the set of all subsequential limits of $\{s_n\}$.

Proposition 2. *Let $\{s_n\}$ be a sequence of real numbers.*

(a) *The lower and upper limits of $\{s_n\}$ always exist in $\bar{\mathbb{R}} = [-\infty, \infty]$ and*

$$\liminf_{n\to\infty} s_n \leq \limsup_{n\to\infty} s_n.$$

(b) *When the ordinary limit $\lim_{n\to\infty} s_n$ exists in $\mathbb{R} = (-\infty, \infty)$,*

$$\liminf_{n\to\infty} s_n = \limsup_{n\to\infty} s_n = \lim_{n\to\infty} s_n.$$

Series: Given an *(infinite) series* $\sum_{k=1}^{\infty} a_k$, we say that the series converges if its partial sum $s_n = \sum_{k=1}^{n} a_k$ converges to s as $n \to \infty$. In this case, we write $s = \sum_{k=1}^{\infty} a_k$.

Proposition 3. $\sum_{k=1}^{\infty} a_k$ *converges if and only if for every $\epsilon > 0$, there is an integer N such that*

$$\left|\sum_{k=n}^{m} a_k\right| \leq \epsilon$$

if $m \geq n \geq N$.

You should also be familiar with some usual tests, such as root and ratio tests, for determining the convergence of $\sum_{k=1}^{\infty} a_k$.

Continuity: Suppose $\mathbb{X}$ and $\mathbb{Y}$ are metric spaces, $E \subset \mathbb{X}$, $p \in E$, and f maps E into $\mathbb{Y}$. Then f is said to be *continuous* at p if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$d_Y\left(f(x), f(p)\right) < \epsilon$$

for all points $x \in E$ for which $d_X(x,p) < \delta$.

Let $f : \mathbb{X} \to \mathbb{Y}$. We say that f is *uniformly continuous* on $\mathbb{X}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$d_Y\left(f(p), f(q)\right) < \epsilon$$

for all p and q in $\mathbb{X}$ for which $d_X(p,q) < \delta$.

Let $f : \mathbb{X} \to \mathbb{Y}$. We say that f is *Lipschitz continuous* if there exists a real constant $K \geq 0$ such that

$$d_Y\left(f(p), f(q)\right) < K \cdot d_X(p,q)$$

for all p and q in $\mathbb{X}$.

Notes: In STAT 512, we mainly deal with the function with range $\mathbb{Y} = \mathbb{R}$ or $\mathbb{R}^n$.

Proposition 4. *f is* Lipschitz continuous $\implies$ *f is uniformly continuous* $\implies$ *f is continuous at every $p \in \mathbb{X}$.*

Derivative of a Real Function: Let $f : [a,b] \to \mathbb{R}$. Define the quotient

$$\phi(t) = \frac{f(t) - f(x)}{t - x} \quad (a < t < b, t \neq x).$$

The derivative of f at x is defined by

$$f'(x) = \lim_{t \to x} \phi(t), \tag{5}$$

provided this limit exists. The higher order derivatives can be defined inductively on f' and its derivatives.

Taylor's Theorem: Suppose f is a real function on $[a,b]$, n is a positive integer, $f^{(n-1)}$ is continuous on $[a,b]$, $f^{(n)}$ exists for every $t \in (a,b)$. Let x, y be distinct points of $[a,b]$. Then, there exists a point θ between x and y such that

$$f(y) = \sum_{k=0}^{n-1} \frac{f^{(k)}(x)}{k!}(y-x)^k + \frac{f^{(n)}(\theta)}{n!}(y-x)^n. \tag{6}$$

See Theorem 5.15 in Rudin [1976] for its proof. For $n = 1$, it reduces to the mean value theorem.

Riemann-Stieltjes Integral: Let α be a monotonically increasing function on $[a,b]$ with $\alpha(a)$ and $\alpha(b)$ being finite. For any real function f which is bounded on $[a,b]$ and a partition P of $[a,b]$ as

$$a = x_0 \leq x_1 \leq \cdots \leq x_{n-1} \leq x_n = b$$

, we write $\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1})$ and

$$U(P,f,\alpha) = \sum_{i=1}^n M_i \Delta\alpha_i \quad \text{ and } \quad L(P,f,\alpha) = \sum_{i=1}^n m_i \Delta\alpha_i,$$

where $M_i = \sup_{x_{i-1} \leq x \leq x_i} f(x)$ and $m_i = \inf_{x_{i-1} \leq x \leq x_i} f(x)$. If $\inf_P U(P,f,\alpha) = \sup_P L(P,f,\alpha)$, then the *Riemann-Stieltjes* integral of f with respect to α over $[a,b]$ exists and is denoted by $\int_a^b f d\alpha \equiv \int_a^b f(x) d\alpha(x)$.

In STAT 512, the Riemann-Stieltjes integral and its notation will appear when we compute the expectation $\int x dF(x)$ of a random variable with cumulative distribution function (CDF) F or its more general statistical functional $\int g(x) dF(x)$ for some function g.

Some Integration Techniques:

Proposition 5. *Let f be an Riemann integrable on $[a, b]$, i.e., $\int_a^b f(x)dx < \infty$. For any $x \in [a, b]$, put*

$$F(x) = \int_a^x f(t)dt.$$

Then F is continuous on $[a, b]$, furthermore, if f is continuous at a point x_0 of $[a, b]$, then F is differentiable at x_0, and

$$F'(x_0) = f(x_0).$$

Proof. See Theorem 6.20 in Rudin [1976] for the proof. □

- **Fundamental theorem of Calculus:** Let f be an Riemann integrable on $[a, b]$. If there is a differentiable function F on $[a, b]$ such that $F' = f$, then

$$\int_a^b f(x)dx = F(b) - F(a).$$

- **Integration by parts:** Suppose F and G are differentiable functions on $[a, b]$, whose derivatives $F' = f$ and $G' = g$ are both Riemannian integrable. Then,

$$\int_a^b F(x)g(x)dx = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)dx.$$

Uniform Convergence of a Sequence of Functions: We say that a sequence of functions $\{f_n\}$ converges *uniformly* on E to a function f if for every $\epsilon > 0$ there is an integer N such that $n \geq N$ implies

$$|f_n(x) - f(x)| \leq \epsilon$$

for all $x \in E$.

Proposition 6. *Let $\{f_n\}$ be a sequence of continuous functions on $[a, b]$.*

(a) If $f_n \to f$ uniformly on $[a, b]$, then f is continuous on $[a, b]$.

(b) Let α be monotonically increasing on $[a, b]$. Suppose each f_n is Riemann-Stieltjes integrable with respect to α. If $f_n \to f$ uniformly on $[a, b]$, then f is also Riemann-Stieltjes integrable on $[a, b]$ and

$$\int_a^b f d\alpha = \lim_{n\to\infty} \int_a^b f_n d\alpha.$$

(c) Suppose that each f_n is differentiable on $[a, b]$ and the numerical sequence $\{f_n(x_0)\}$ converges for some point x_0 on $[a, b]$. If $\{f_n'\}$ converges uniformly on $[a, b]$, then $\{f_n\}$ converges uniformly on $[a, b]$ to a function f and

$$f'(x) = \lim_{n\to\infty} f_n'(x) \quad \text{for any } x \in [a, b].$$

Proof. See the proof in Theorem 7.11, 7.16, 7.17 in Rudin [1976]. □

Differentiation of a Multivariate Function: Suppose E is an open set in $\mathbb{R}^n$, $\boldsymbol{f} : E \to \mathbb{R}^m$, and $\boldsymbol{x} \in E$. If there exists a linear transformation A of $\mathbb{R}^n$ into $\mathbb{R}^m$ such that

$$\lim_{\boldsymbol{h}\to 0} \frac{|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x}) - A\boldsymbol{h}|}{|\boldsymbol{h}|} = 0,$$

then we say that $\boldsymbol{f}$ is *differentiable* at $\boldsymbol{x}$, and we write $\boldsymbol{f}'(\boldsymbol{x}) = A$. If $\boldsymbol{f}$ is differentiable at every $\boldsymbol{x} \in E$, we say that $\boldsymbol{f}$ is *differentiable* in E.
When $m = 1$, $A = \nabla f(\boldsymbol{x})$ becomes the gradient of f.

Partial Derivatives: We again consider a function $\boldsymbol{f}$ that maps an open set $E \subset \mathbb{R}^n$ into $\mathbb{R}^m$. Let $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_n\}$ and $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_m\}$ be the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. The *components* of $\boldsymbol{f}$ are the real functions $f_1, ..., f_m$ defined by

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=1}^{m} f_i(\boldsymbol{x})\boldsymbol{u}_i \quad (\boldsymbol{x} \in E).$$

For $\boldsymbol{x} \in E$, $1 \leq i \leq m$, $1 \leq j \leq n$, we define the partial derivative as

$$\frac{\partial f_i}{\partial x_j}(\boldsymbol{x}) = \lim_{t \to 0} \frac{f_i(\boldsymbol{x} + t\boldsymbol{e}_j) - f_i(\boldsymbol{x})}{t},$$

provided the limit exists.

Inverse Function Theorem: Suppose $\boldsymbol{f}$ is a continuously differentiable (meaning that all of its first-order partial derivatives are continuous) mapping of an open set $E \subset \mathbb{R}^n$ into $\mathbb{R}^n$, $\boldsymbol{f}'(\boldsymbol{a})$ is invertible for some $\boldsymbol{a} \in E$, and $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$. Then

(a) there exist open sets U and V in $\mathbb{R}^n$ such that $\boldsymbol{a} \in U, \boldsymbol{b} \in V$, $\boldsymbol{f}$ is one-to-one on U, and $\boldsymbol{f}(U) = V$.

(b) if $\boldsymbol{g}$ is the inverse of $\boldsymbol{f}$, which exists by (a), defined in V by

$$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{x} \quad (\boldsymbol{x} \in U),$$

then $\boldsymbol{g}$ is continuously differentiable on V.

See the proof in Theorem 9.24 in Rudin [1976].

Implicit Function Theorem: Let $\boldsymbol{f}$ be a continuously differentiable mapping of an open set $E \subset \mathbb{R}^{n+m}$ into $\mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{a}, \boldsymbol{b}) = 0$ for some point $(\boldsymbol{a}, \boldsymbol{b}) \in E$.
Put $A = (A_{\boldsymbol{x}}, A_{\boldsymbol{y}}) = \boldsymbol{f}'(\boldsymbol{a}, \boldsymbol{y})$, where A is a linear transformation from $\mathbb{R}^{n+m}$ to $\mathbb{R}^n$, $A_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$ is the linear transformation of its first n coordinates, and $A_{\boldsymbol{y}} : \mathbb{R}^m \to \mathbb{R}^n$ is the linear transformation of its last m coordinates. We assume that $A_{\boldsymbol{x}}$ is invertible.
Then, there exist open sets $U \subset \mathbb{R}^{n+m}$ and $W \subset \mathbb{R}^m$, with $(\boldsymbol{a}, \boldsymbol{b}) \in U$ and $\boldsymbol{b} \in W$, having the following property:

(a) To every $\boldsymbol{y} \in W$ corresponds a unique $\boldsymbol{x}$ such that

$$(\boldsymbol{x}, \boldsymbol{y}) \in U \quad \text{and} \quad \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}.$$

(b) If the above $\boldsymbol{x}$ is defined to be $\boldsymbol{g}(\boldsymbol{y})$, then $\boldsymbol{g}$ is a continuously differentiable mapping of W into $\mathbb{R}^n$, $\boldsymbol{g}(\boldsymbol{b}) = \boldsymbol{a}$,

$$\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y}), \boldsymbol{y}) = \boldsymbol{0} \quad (\boldsymbol{y} \in W) \quad \text{and} \quad \boldsymbol{g}'(\boldsymbol{b}) = -(A_{\boldsymbol{x}})^{-1} A_{\boldsymbol{y}}.$$

See the proof in Theorem 9.28 in Rudin [1976].

Differentiation of Integrals: Suppose

(a) $\varphi(x, t)$ is defined for $a \leq x \leq b, c \leq t \leq d$;

(b) α is a non-decreasing function on $[a, b]$;

(c) $\varphi(\cdot, t)$ is Riemann-Stieltjes integrable with respect to α for every $t \in [c, d]$;

(d) $\frac{\partial \varphi}{\partial t}(x, \cdot)$ is continuous on (c, d) for all $x \in [a, b]$.

Define $f(t) = \int_a^b \varphi(x, t)\, d\alpha(x)$ with $t \in [c, d]$. Then, $\frac{\partial \varphi}{\partial t}(\cdot, s)$ is Riemann-Stieltjes integrable with respect to α for every $s \in [c, d]$, $f'(s)$ exists, and

$$f'(s) = \int_a^b \frac{\partial \varphi}{\partial t}(x, s)\, d\alpha(x).$$

See Theorem 9.42 in Rudin [1976] for one possible proof.

In STAT 512, we sometimes need to consider the interchanges of differentiation/limit and integration with measurable functions defined on a more general measure space $(\Omega, \mathcal{F}, \mu)$; see Remark 1 below. Therefore, we provide some sufficient conditions under which the interchanges are valid.

Theorem 7. *Let $\{f_n\}$ be a sequence of (Borel) measurable functions on $(\Omega, \mathcal{F}, \mu)$.*

(a) (Fatou's Lemma). *If $f_n \geq 0$, then*

$$\int \liminf_{n\to\infty} f_n\, d\mu \leq \liminf_{n\to\infty} \int f_n\, d\mu.$$

(b) (Monotone Convergence Theorem). *If $0 \leq f_1 \leq f_2 \leq \cdots$ and $\lim_{n\to\infty} f_n = f$ almost everywhere, then*

$$\int \lim_{n\to\infty} f_n\, d\mu = \lim_{n\to\infty} \int f_n\, d\mu.$$

(c) (Dominated Convergence Theorem). *If $\lim_{n\to\infty} f_n = f$ almost everywhere and there exists an integrable function g (*i.e., $\int g d\mu < \infty$*) such that $|f_n| \leq g$ almost everywhere, then*

$$\int \lim_{n\to\infty} f_n\, d\mu = \lim_{n\to\infty} \int f_n\, d\mu.$$

(d) (Interchange of Differentiation and Integration). *For any fixed $\theta \in \mathbb{R}$, let $f(\omega, \theta)$ be a (Borel) measurable function on Ω. Suppose that $\frac{\partial f(\omega,\theta)}{\partial \theta}$ exists almost everywhere for $\theta \in (a, b) \subset \mathbb{R}$ and that $\left|\frac{\partial f(\omega,\theta)}{\partial \theta}\right| \leq g(\omega)$ almost everywhere, where $g(\omega)$ is an integrable function on Ω. Then, for each $\theta \in (a, b)$, $\frac{\partial f(\omega,\theta)}{\partial \theta}$ is integrable and*

$$\frac{d}{d\theta} \int f(\omega, \theta)\, d\mu = \int \frac{\partial f(\omega, \theta)}{\partial \theta}\, d\mu.$$

Proof. See Theorem 1.5.5, 1.5.7, 1.5.8 in Durrett [2019] as well as Theorem 1.1 and Example 1.8 in Shao [2003]. ☐

Leibniz Integral Rule: Let $f(x, t)$ be a function such that both $f(x, t)$ and its partial derivative $\frac{\partial f}{\partial t}(x, t)$ are continuous with respect to both t and x in some open set $U \subset \mathbb{R}^2$, including $a(x) \leq t \leq b(x)$ and $x_0 \leq x \leq x_1$. Suppose also that the functions $a(x)$ and $b(x)$ are both continuously differentiable for $x_0 \leq x \leq x_1$. Then, for any $x_0 \leq x \leq x_1$,

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x, t) dt\right) = f(x, b(x)) \cdot b'(x) - f(x, a(x)) \cdot a'(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x, t) dt.$$

Another approach[1] to proving the above two results (differentiation of integrals or Leibniz integral rules) is to use the Fubini's theorem:

[1] See `https://en.wikipedia.org/wiki/Leibniz_integral_rule`.

Theorem 8 (Fubini)**.** *Suppose X and Y are σ-finite measure spaces and $X\times Y$ is given the product measure. If f is $X\times Y$ integrable, meaning that f is a measurable function and $\int_{X\times Y}|f(\boldsymbol{x},\boldsymbol{y})|d(\boldsymbol{x},\boldsymbol{y})<\infty$, then*

$$\int_X\left(\int_Y f(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y}\right)d\boldsymbol{x}=\int_Y\left(\int_X f(\boldsymbol{x},\boldsymbol{y})\right)d\boldsymbol{y}=\int_{X\times Y} f(\boldsymbol{x},\boldsymbol{y})d(\boldsymbol{x},\boldsymbol{y}).$$

Remark 1. *A* measure *$\mu:\mathcal{F}\to\mathbb{R}$ is a nonnegatice countably additive set function with $\mathcal{F}$ being the σ-field and*

(a) $\mu(A)\geq\mu(\emptyset)=0$ for all $A\in\mathcal{F}$, and

(b) if $A_n\in\mathcal{F}$ is a countable sequence of disjoint sets, then

$$\mu(\cup_i A_n)=\sum_i\mu(A_n).$$

*Let Ω be the sample space. Then, the measure μ is said to be **σ-finite** if there is a sequence of sets $A_n\in\mathcal{F}$ so that $\mu(A_n)<\infty$ and $\cup_n A_n=\Omega$.*

- *The probability measure $\mathbb{P}$ is a σ-finite measure with $\mathbb{P}(\Omega)=1$.*
- *The Lebesgue measure on $\mathbb{R}^n$ is also a σ-finite measure. When $X\times Y$ are measurable in $\mathbb{R}^{n+m}$, the Fubini's theorem provides a feasible way to compute double/multiple integrals.*

Gamma Function and Stirling's Formula: For $x\in(0,\infty)$, the Gamma function is $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$. While the exact value of $\Gamma(x+1)$ is intractable for some $x\in(0,\infty)$, one can approximate $\Gamma(x+1)$ when x is large by *Stirling's formula* (see 8.22 in Rudin 1976)

$$\lim_{x\to\infty}\frac{\Gamma(x+1)}{(x/e)^x\sqrt{2\pi x}}=1. \tag{7}$$

2 Linear Algebra

The review of Linear Algebra will be conducted on Topic 8 during the regular lectures. Other useful references for linear algebra include

- S. Axler. *Linear algebra done right.* Springer, 3 edition, 2015. [Axler, 2015]
- R. A. Horn and C. R. Johnson. Matrix analysis. Cambridge university press, 2 edition, 2012. [Horn and Johnson, 2012]

3 Some Inequalities

Cauchy-Schwarz Inequality: For all vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ if an inner product space, the *Cauchy-Schwarz inequality* can be stated as

$$|\langle\boldsymbol{x},\boldsymbol{y}\rangle|\leq||\boldsymbol{x}||\cdot||\boldsymbol{y}||, \tag{8}$$

where $||\boldsymbol{x}||=\sqrt{\langle\boldsymbol{x},\boldsymbol{x}\rangle}$. In particular, equality hold if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent. In $\mathbb{R}^n$ with the standard inner product, the Cauchy-Schwarz inequality becomes:

$$\left(\sum_{i=1}^n x_iy_i\right)^2\leq\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right).$$

Hölder Inequality: Let S be a measurable subset of $\mathbb{R}^n$ with the Lebesgue measure. Given two measurable functions[2] on S, the *Hölder inequality* is

$$\int_S |f(x)g(x)|dx \leq \left(\int_S |f(x)|^p dx\right)^{\frac{1}{p}} \left(\int_S |g(x)|^q dx\right)^{\frac{1}{q}}, \tag{9}$$

where $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. When $p = \infty$, $\left(\int_S |f(x)|^p dx\right)^{\frac{1}{p}} := ||f||_\infty$ stands for the essential supremum[3] of $|f|$. If $p, q \in (1, \infty)$, then the equality holds if and only if there exist real numbers $\alpha, \beta \geq 0$, not both of them zero, such that $\alpha|f(x)|^p = \beta|g(x)|^q$ almost everywhere on S with respect to the Lebesgue measure.

When $S = \{1, ..., n\}$ with the counting measure, the Hölder inequality becomes

$$\sum_{i=1}^n |x_i y_i| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^n |y_i|^q\right)^{\frac{1}{q}} \qquad \text{for any } (x_1, ..., x_n), (y_1, ..., y_n) \in \mathbb{R}^n.$$

Minkowski Inequality: Let S be a measurable subset of $\mathbb{R}^n$. Given two measurable functions f and g with $||f||_p \equiv \left(\int_S |f(x)|^p dx\right)^{\frac{1}{p}} < \infty, ||g||_p < \infty$, *Minkowski inequality* can be stated as

$$||f + g||_p \leq ||f||_p + ||g||_p\,, \tag{10}$$

with equality for $p \in [1, \infty)$ if and only if $f = \lambda g$ for some $\lambda \geq 0$ or $g \equiv 0$. In the case when $p = \infty$, the Minkowski inequality is still valid and $||f||_\infty$ is the essential supremum of $|f|$.

Like Hölder inequality, the Minkowski inequality can be specialized to sequences and vectors via the counting measure

$$\left(\sum_{i=1}^n |x_k + y_k|^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^n |x_k|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |y_k|^p\right)^{\frac{1}{p}}$$

for all real numbers $x_1, ..., x_n, y_1, ..., y_n$, where n is the cardinality of S.

Jensen's Inequality: For a real *convex* function φ, real numbers $x_1, ..., x_k$ in its domain, and positive weights $a_1, ..., a_k$, *Jensen's inequality* can be stated as

$$\varphi\left(\frac{\sum_{i=1}^k a_i x_i}{\sum_{i=1}^k a_i}\right) \leq \frac{\sum_{i=1}^k a_i \varphi(x_i)}{\sum_{i=1}^k a_i}. \tag{11}$$

Equality holds if and only if $x_1 = \cdots x_k$ or φ is linear on a domain containing $x_1, ..., x_k$.

3.1 Probability Form

We also present the above inequalities under the context of probability theory. For the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathbb{E}$ denote the expectation operator. Given two random variables X and Y, we have that

- **Cauchy-Schwarz inequality:**

  $$|\mathbb{E}_{X,Y}(XY)|^2 \leq \mathbb{E}_X(X^2) \cdot \mathbb{E}_Y(Y^2),$$

  where equality holds if and only if either $\mathbb{E}_X(X^2) = 0$ or $\mathbb{E}_Y(Y^2) = 0$, or $\mathbb{P}_{X,Y}(X = cY) = 1$ for some nonzero constant $c \in \mathbb{R}$.
  A useful corollary of the Cauchy-Schwarz inequality is that

  $$|\text{Cov}(X, Y)|^2 \leq \text{Var}(X)\text{Var}(Y),$$

[2] https://en.wikipedia.org/wiki/Measurable_function.
[3] https://en.wikipedia.org/wiki/Essential_infimum_and_essential_supremum.

where $\text{Cov}(X,Y) = \mathbb{E}_{X,Y}\left[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right]$ is the covariance between X and Y and $\text{Var}(X) = \mathbb{E}_X\left[(X-\mathbb{E}(X))^2\right]$ is the variance of X.

- **Hölder inequality:**
$$\mathbb{E}_{X,Y}|XY| \le \left(\mathbb{E}_X|X|^p\right)^{\frac{1}{p}}\left(\mathbb{E}_Y|Y|^q\right)^{\frac{1}{q}} \equiv ||X||_p\,||Y||_q$$
with $p,q \in [1,\infty]$ and $\frac{1}{p}+\frac{1}{q}=1$, where equality holds if and only if $\mathbb{P}_{X,Y}\left(|X|^p = c|Y|^q\right)=1$ for some nonzero constant c. Specifically, when $p=\infty$, $||X||_\infty = \inf\left\{M : \mathbb{P}_X(|X|>M)=0\right\}$.
Let $1 \le r < s < \infty$ and define $p=\frac{s}{r}, q=\frac{p}{p-1}$. Applying Hölder inequality to the random variables $|X|^r$ and $\mathbb{1}_\Omega$ yields that
$$\mathbb{E}_X\left[|X|^r\right] \le \left(\mathbb{E}_X\left[|X|^s\right]\right)^{\frac{r}{s}}.$$
It demonstrates that the r-th absolute moment is finite whenever the (higher) s-th moment is finite. (This can also be proved using Jensen's inequality.)

- **Minkowski Inequality:**
$$\left[\mathbb{E}_{X,Y}|X+Y|^p\right]^{\frac{1}{p}} \le \left[\mathbb{E}_X|X|^p\right]^{\frac{1}{p}} + \left[\mathbb{E}_Y|Y|^p\right]^{\frac{1}{p}}$$
for $p \in [1,\infty)$, where equality holds if and only if $\mathbb{P}_{X,Y}\left(X=cY\right)=1$ for some nonzero constant c or $\mathbb{P}_Y(Y=0)=1$.

- **Jensen's Inequality:** Given a convex function φ,
$$\varphi\left(\mathbb{E}_X(X)\right) \le \mathbb{E}_X\left[\varphi(X)\right],$$
where equality holds if and only if either $\mathbb{P}_X(X=c)=1$ for some constant c, or for every line $a+bx$ that is tangent to φ at $\mathbb{E}_X[X]$, $\mathbb{P}_X\left(\varphi(x)=a+bx\right)=1$.

Remark 2. *Jensen's inequality provides an insight into the rationale behind the maximum likelihood estimator (MLE): Suppose $X_1,...,X_n$ are independent and identically distributed (i.i.d.) from a distribution with density $p(x;\theta_0)$, where $\theta_0 \in \Theta$ and Θ is a bounded subset in $\mathbb{R}^n$. We find the MLE by maximizing the log-likelihood function $\ell(\theta)=\sum_{i=1}^n \log p(X_i;\theta)$ because by Jensen's inequality,*
$$\begin{aligned}
\mathbb{E}_{\theta_0}[\ell(\theta_0)] - \mathbb{E}_{\theta_0}[\ell(\theta)] &= n\cdot\mathbb{E}_{\theta_0}\left[\log\frac{p(X_1;\theta_0)}{p(X_1;\theta)}\right] \\
&= n\cdot\mathbb{E}_{\theta_0}\left[-\log\frac{p(X_1;\theta)}{p(X_1;\theta_0)}\right] \\
&\ge n\cdot\left(-\log\mathbb{E}_{\theta_0}\left[\frac{p(X_1;\theta)}{p(X_1;\theta_0)}\right]\right) \\
&= -n\log\int\frac{p(x;\theta)}{p(x;\theta_0)}\cdot p(x;\theta_0)dx \\
&= 0.
\end{aligned}$$
Also, in STAT 513, we will leverage the Jensen's inequality to prove the non-descending properties of EM algorithm [Dempster et al., 1977, Wu, 1983, McLachlan and Krishnan, 2007].

The proof of these inequalities under the context of probability theory can be found in https://www.math.mcgill.ca/dstephens/556-2014/Handouts/Math556-05-Inequalities.pdf.

References

S. Axler. *Linear algebra done right.* Springer, 3 edition, 2015.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

R. Durrett. *Probability: Theory and Examples*, volume 49. Cambridge university press, 5th edition, 2019.

R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge university press, 2 edition, 2012.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* John Wiley & Sons, 2007.

W. Rudin. *Principles of mathematical analysis.* McGraw-hill New York, 3 edition, 1976.

J. Shao. *Mathematical statistics.* Springer Science & Business Media, 2th edition, 2003.

C. J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.