Theoretical Statistics, STOR 655

Backgound and Preliminary Material

Andrew Nobel

January 2023

Order, Minima, and Maxima

Multiplication and Addition

Recall: For any numbers a, b

(1) If $a, b \geq 0$ or $a, b \leq 0$ then $ab \geq 0$

(2) If $a \geq 0$ and $b \leq 0$ or vice-versa then $ab \leq 0$

(3) If $a, b \geq 0$ then $a + b \geq 0$

(4) If $a, b \leq 0$ then $a + b \leq 0$.

Note: (1)-(4) continue to hold if we replace $\leq$ and $\geq$ by $<$ and $>$, respectively

The Usual Order Relation

Definition: For $a, b \in \mathbb{R}$ write $a \leq b$ if $(b - a) \geq 0$ and $a < b$ if $(b - a) > 0$

Basic Properties

1. If $a \leq b$ and $b \leq a$ then $a = b$
2. If $a \leq b$ then $-b \leq -a$
3. If $a \leq b$ and $c \leq d$ then $a + c \leq b + d$
4. If $0 \leq a \leq b$ and $0 \leq c \leq d$ then $ac \leq bd$

Note: (2)-(4) continue to hold if we replace $\leq$ by $<$

Maxima and Minima of Finite Sequences

Definition: Let $a_1, \ldots, a_n \in \mathbb{R}$

- $\max\{a_1, \ldots, a_n\}$ is any element a_j such that $a_i \leq a_j$ for $i = 1, \ldots, n$
- $\min\{a_1, \ldots, a_n\}$ is any element a_j such that $a_i \geq a_j$ for $i = 1, \ldots, n$

Other Notation

- $\max_{1 \leq i \leq n} a_i$ or simply $\max_i a_i$
- $\min_{1 \leq i \leq n} a_i$ or simply $\min_i a_i$

Maxima and Minima, cont.

Basic Properties: Let $a_1, \ldots, a_n \in \mathbb{R}$ and $b_1, \ldots, b_n \in \mathbb{R}$ be finite sequences

1. If $a_i \leq b_i$ for each i, then $\max_i a_i \leq \max_i b_i$ and $\min_i a_i \leq \min_i b_i$
2. $\min_i a_i \leq a_j \leq \max_i a_i$ for $j = 1, \ldots, n$
3. $-\min_i a_i = \max_i(-a_i)$ and $-\max_i a_i = \min_i(-a_i)$
4. If $c \geq 0$ and b are constants then $c \max_i a_i + b = \max_i(c\, a_i + b)$
5. $\max_i(a_i + b_i) \leq \max_i a_i + \max_i b_i$
6. $\min_i(a_i + b_i) \geq \min_i a_i + \min_i b_i$
7. $\max_i a_i - \max_i b_i \leq \max_i |a_i - b_i|$

Suprema and Infima

Definition: Let $A \subseteq \mathbb{R}$ be bounded. Recall that

- $\sup(A)$ = least upper bound for A
- $\inf(A)$ = greatest lower bound for A

Existence of $\sup$ and $\inf$ follows from construction of the real numbers.

Basic Properties and Conventions

1. If A is not bounded, then $\sup(A) = +\infty$ or $\inf(A) = -\infty$, or both
2. By convention $\sup(\emptyset) = -\infty$ and $\inf(\emptyset) = +\infty$
3. If $A \subseteq B$ then $\sup(A) \leq \sup(B)$ while $\inf(A) \geq \inf(B)$

Order Relations for Maxima and Minima of Functions

Fact: Let $f, g : \mathcal{X} \to \mathbb{R}$ be functions.

(1) $\inf_{x \in \mathcal{X}} f(x) \le f(x_0) \le \sup_{x \in \mathcal{X}} f(x)$ for every $x_0 \in \mathcal{X}$

(2) $-\sup_{x \in \mathcal{X}} f(x) = \inf_{x \in \mathcal{X}}(-f(x))$

(3) $\sup_{x \in \mathcal{X}}\{f(x) + g(x)\} \le \sup_{x \in \mathcal{X}} f(x) + \sup_{x \in \mathcal{X}} g(x)$

(4) If $\mathcal{X}_0 \subseteq \mathcal{X}$ then $\sup_{x \in \mathcal{X}_0} f(x) \le \sup_{x \in \mathcal{X}} f(x)$

Fact: If $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is any function

$$\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} h(x, y) \le \inf_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} h(x, y)$$

Argmax and Argmin

Definition: The *argmax* of a function $f : \mathcal{X} \to \mathbb{R}$ is the set of points $y \in \mathcal{X}$ where f is maximized

$$\begin{aligned} \operatorname*{argmax}_{x \in \mathcal{X}} f(x) &= \{y \in \mathcal{X} : f(y) \geq f(x) \text{ for all } x \in \mathcal{X}\} \\ &= \left\{ y \in \mathcal{X} : f(y) = \max_{x \in \mathcal{X}} f(x) \right\} \end{aligned}$$

Similarly, the *argmin* of f is the set of points $y \in \mathcal{X}$ where f is minimized

$$\begin{aligned} \operatorname*{argmin}_{x \in \mathcal{X}} f(x) &= \{y \in \mathcal{X} : f(y) \leq f(x) \text{ for all } x \in \mathcal{X}\} \\ &= \left\{ y \in \mathcal{X} : f(y) = \min_{x \in \mathcal{X}} f(x) \right\} \end{aligned}$$

Argmax and Argmin, cont.

Note that $\mathrm{argmax}_{x \in \mathcal{X}} f(x)$ is a subset of $\mathcal{X}$

- $\max_{x \in \mathcal{X}} f(x)$ is the maximum value of $f(x)$ if this exists
- $\mathrm{argmax}_{x \in \mathcal{X}} f(x)$ is the set of arguments x achieving the maximum value
- $\mathrm{argmax}_{x \in \mathcal{X}} f(x)$ is non-empty iff $\max_{x \in \mathcal{X}} f(x)$ defined

Note that $\mathrm{argmin}_{x \in \mathcal{X}} f(x)$ is a subset of $\mathcal{X}$

- $\min_{x \in \mathcal{X}} f(x)$ is the minimum value of $f(x)$ if this exists
- $\mathrm{argmin}_{x \in \mathcal{X}} f(x)$ is the set of arguments x achieving the minimum value
- $\mathrm{argmin}_{x \in \mathcal{X}} f(x)$ is non-empty iff $\min_{x \in \mathcal{X}} f(x)$ defined

Matrix Algebra

Inner Product

Definition: The *inner product* of two vectors $u, v \in \mathbb{R}^d$ is given by

$$\langle u, v \rangle \; = \; u^t v \; = \; \sum_{i=1}^{d} u_i \, v_i$$

Basic Properties: Let $u, v, w \in \mathbb{R}^d$ and $a, b \in \mathbb{R}$

1. $\langle u, v \rangle = \langle v, u \rangle$
2. $\langle au, bv \rangle = ab \, \langle u, v \rangle$
3. $\langle u + w, v \rangle = \langle u, v \rangle + \langle w, v \rangle$

Euclidean Norm

Definition: The *Euclidean norm* of a vector $u \in \mathbb{R}^d$ is

$$||u|| = \langle u, u \rangle^{1/2} = (u_1^2 + \cdots + u_d^2)^{1/2}$$

Basic Properties

1. $||u|| \geq 0$ with equality if and only if $u = 0$
2. For $a \in \mathbb{R}$, $||a\, u|| = |a|\, ||u||$
3. $||u + v||^2 = ||u||^2 + 2\langle u, v \rangle + ||v||^2$
4. $|\langle u, v \rangle| = |u^t v| \ \leq \ ||u||\, ||v||$ (Cauchy-Schwarz inequality)
5. $||u + v|| \ \leq \ ||u|| + ||v||$ (triangle inequality)
6. $|\, ||u|| - ||v||\, | \ \leq \ ||u - v||$ (reverse triangle inequality)

Orthogonality and Projections

Definition: Vectors $u, v \in \mathbb{R}^n$ are orthogonal, written $u \perp v$, if $\langle u, v \rangle = 0$

Defn: Let V be a subspace of $\mathbb{R}^n$. The *projection* of $u \in \mathbb{R}^n$ onto V is the vector $w \in V$ closest to u. Formally,

$$\text{proj}_V(u) = \operatorname*{argmin}_{w \in V} ||u - w||$$

Fact: Let $V = \{\alpha v : \alpha \in \mathbb{R}\}$ be the 1-d subspace generated by $v \in \mathbb{R}^n$

1. $\text{proj}_V(u) = \langle u, v \rangle \, v / ||v||^2$
2. $(u - \text{proj}_V(u)) \perp v$

Orthogonal Matrices

Vectors $u_1, \ldots, u_n$ are *orthonormal* if $\langle u_i, u_j \rangle = \mathbb{I}(i = j)$ for $1 \leq i, j \leq n$

A matrix $A \in \mathbb{R}^{n \times n}$ is *orthogonal* if $A^t A = I$. If A is orthogonal then

- $A^{-1} = A^t$
- $A\, A^t = I$
- the rows and columns of A are orthonormal
- the eigenvalues $\lambda_i(A) \in \{+1, -1\}$
- $\det(A) \in \{+1, -1\}$

Quadratic Forms

Each symmetric matrix $A \in \mathbb{R}^{n \times n}$ has an associated *quadratic form* $q_A : \mathbb{R}^n \to \mathbb{R}$ defined by

$$q_A(u) = u^t A\, u = \sum_{i=1}^{n} \sum_{j=1}^{n} u_i\, a_{ij}\, u_j$$

- A is *non-negative definite* ($A \geq 0$) if $u^t A\, u \geq 0$ for every u
- A is *positive definite* ($A > 0$) if $u^t A\, u > 0$ for every $u \neq 0$

Fact: Let A $n \times n$ be symmetric.

- $A \geq 0$ iff all its eigenvalues are non-negative
- $A > 0$ iff all its eigenvalues are positive

Trace of a Matrix

Definition: The *trace* of a matrix $A \in \mathbb{R}^{n \times n}$ is the sum of its diagonal elements

$$\text{tr}(A) \; = \; \sum_{i=1}^{n} a_{ii}$$

- $\text{tr}(A) =$ sum of eigenvalues of A
- $\text{tr}(A) = \text{tr}(A^t)$
- If B is $n \times n$ then $\text{tr}(AB) = \text{tr}(BA)$

Frobenius Norm

Definition: The *Frobenius norm* of a matrix $A \in \mathbb{R}^{m \times n}$ is

$$||A|| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

Basic Properties

- $||A||^2 = \text{tr}(A^t A)$
- $||A|| = 0$ if and only if $A = 0$
- $||bA|| = |b| \, ||A||$
- $||A + B|| \leq ||A|| + ||B||$
- $||AB|| \leq ||A|| \, ||B||$

Rank of a Matrix

Definition: Let $A \in \mathbb{R}^{m \times n}$ be an m x n matrix

- row-space of $A =$ span of the rows of A (subspace of $\mathbb{R}^n$)
- col-space of $A =$ span of the cols of A (subspace of $\mathbb{R}^m$)
- row-rank(A) $:=$ dim of the row-space of A (at most n)
- col-rank(A) $:=$ dim of the col-space of A (at most m)

Fact: row-rank$(A) =$ col-rank(A)

Definition: The *rank* of A is the common value of the row and column ranks

Basic Properties of the Rank

- If $A \in \mathbb{R}^{m \times n}$ then $\text{rank}(A) \leq \min\{m, n\}$
- $\text{rank}(A\,B) \leq \min\{\text{rank}(A), \text{rank}(B)\}$
- $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$
- $\text{rank}(A) = \text{rank}(A^t) = \text{rank}(A^t A) = \text{rank}(AA^t)$
- $A \in \mathbb{R}^{n \times n}$ has at most $\text{rank}(A)$ non-zero eigenvalues
- $A \in \mathbb{R}^{n \times n}$ is invertible iff $\text{rank}(A) = n$, that is, A is of full rank

Outer Products

Definition: The *outer product* uv^t of vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ is an $m \times n$ matrix with entries

$$(uv^t)_{ij} = u_i v_j$$

- If $u, v \neq 0$ then rank$(uv^t) = 1$
- $||uv^t|| = ||u|| \, ||v||$
- If $m = n$ then tr$(uv^t) = \langle u, v \rangle$

The Spectral Theorem

Spectral Theorem: If $A \in \mathbb{R}^{n \times n}$ is symmetric there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A

Corollary: If $A \in \mathbb{R}^{n \times n}$ is symmetric then it can be expressed in the form

$$A \; = \; \Gamma \, D \, \Gamma^t$$

where $\Gamma \in \mathbb{R}^{n \times n}$ is orthogonal and $D = \text{diag}(\lambda_1(A), \ldots, \lambda_n(A))$ has the eigenvalues of A on the diagonal, with all other values equal to zero

- $A^k = \Gamma D^k \Gamma^t$ for $k \geq 1$
- If $A \geq 0$ we may define $A^\alpha = \Gamma D^\alpha \Gamma^t$ for $\alpha > 0$

Courant Fischer Theorem

Thm: Let $A \in \mathbb{R}^{n \times n}$ be symmetric with eigenvalues $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$.

$$\lambda_1(A) = \max_{v \neq 0} \frac{v^t A v}{v^t v} = \max_{v:||v||=1} v^t A v$$

$$\lambda_n(A) = \min_{v \neq 0} \frac{v^t A v}{v^t v} = \min_{v:||v||=1} v^t A v$$

$$\lambda_i(A) = \max_{V:\dim(V)=i} \min_{v \in V, ||v||=1} v^t A v$$

Continuous Functions and Compact Sets

Continuous Functions

Definition: Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. We say that f is

1. *bounded* if there exists $M < \infty$ such that $|f(x)| \leq M$ for all x.
2. *continuous at* $x \in \mathbb{R}^d$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that $||x - y|| < \delta$ implies $|f(x) - f(y)| < \epsilon$
3. *continuous* if it is continuous at every $x \in \mathbb{R}^d$
4. *uniformly continuous* if for every $\epsilon > 0$ there exists $\delta > 0$ such that $||x - y|| < \delta$ implies $|f(x) - f(y)| < \epsilon$

Distinction

- Continuity: δ depends on ϵ *and* x
- Uniformly continuity: δ depends only on ϵ

Continuous Functions, cont.

Fact: A set $K \subseteq \mathbb{R}^d$ is compact iff it is closed and bounded

Fact: If $K \subseteq \mathbb{R}^d$ is compact and $f : K \to \mathbb{R}$ is continuous, then f is uniformly continuous and bounded on K

Definition: The *support* of a function $f : \mathbb{R}^d \to \mathbb{R}$ is

$$\text{supp}(f) \; = \; \overline{\{x : f(x) \neq 0\}}$$

Note: supp(f) is closed by definition, and compact if it is bounded

Continuous Functions, cont.

Definition

1. $C_b(\mathbb{R}^d) =$ family of bounded continuous functions $f : \mathbb{R}^d \to \mathbb{R}$
2. $C_o(\mathbb{R}^d) =$ family of continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ with compact support

Fact

1. Every $f \in C_o(\mathbb{R}^d)$ is uniformly continuous
2. $C_o(\mathbb{R}^d) \subseteq C_b(\mathbb{R}^d)$
3. Every $f \in C_b(\mathbb{R}^d)$ is Borel measurable

Multivariate Calculus

Multivariate Differentiation: Total Derivative

Definition: A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is *differentiable* at $x \in \mathbb{R}^d$ if there exists a matrix $A \in \mathbb{R}^{k \times d}$ such that

$$\lim_{h \to 0} \frac{||f(x+h) - f(x) - Ah||}{||h||} \; = \; 0$$

which can be written in the equivalent form

$$f(x+h) \; = \; f(x) + Ah + o(||h||)$$

The (unique) matrix A satisfying these conditions is called the *total derivative* of f at x, and denoted by $Df(x)$ or $\dot{f}(x)$

Total Derivatives

First Examples: Consider a function $f : \mathbb{R}^d \to \mathbb{R}^k$

- If $d = k = 1$ then $Df(x) = f'(x)$ coincides with ordinary derivative
- If $f(x) = c$ is constant then $Df(x) = \mathbf{0}$ is the $k \times d$ zero matrix
- If $f(x) = Bx$ is linear then $Df(x) = B$
- If $f(x) = x^t V x$ where $V \in \mathbb{R}^{d \times d}$ is symmetric then $DF(x) = 2x^t V$

Chain Rule: If $f : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at x and $g : \mathbb{R}^k \to \mathbb{R}^l$ is differentiable at $f(x)$, then $g \circ f$ is differentiable at x and

$$D(g \circ f)(x) \ = \ Dg(f(x))\, Df(x)$$

Jacobians

Note that $f : \mathbb{R}^d \to \mathbb{R}^k$ can be written $f = (f_1, \ldots, f_k)$ where $f_i : \mathbb{R}^d \to \mathbb{R}$

Definition: The *Jacobian* of f at x is the $k \times d$ matrix of partial derivatives

$$J_f(x) \ = \ \left[\frac{\partial f_i}{\partial x_j}(x) : 1 \leq i \leq k, 1 \leq j \leq d\right]$$

Fact: Jacobians and Total Derivatives

(a) If f is differentiable at x then $J_f(x)$ exists and is equal to $Df(x)$

(b) If the Jacobian J_f exists and is continuous at x then f is differentiable at x and $J_f(x) = Df(x)$

Gradients and Hessians

Definition. Let $f : \mathbb{R}^d \to \mathbb{R}$

- The *gradient* of f at x is the $d \times 1$ vector of partial derivatives

$$\nabla f(x) \;=\; \left(\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_d}(x) \right)^t$$

When derivative $Df(x)$ exists, $\nabla f(x)$ exists and is equal to $Df(x)^t$

- The *Hessian* of f at x is the $d \times d$ matrix of second partial derivatives

$$\nabla^2 f(x) \;=\; \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x) : 1 \leq i, j \leq d \right]$$

If the second partials are continuous, then $\nabla^2 f(x)$ is symmetric

Multivariate Taylor's Theorem I

Fact: If $f : \mathbb{R}^d \to \mathbb{R}^k$ has continuous partial derivatives $\partial f_i / \partial x_j$ at each point in $\mathbb{R}^d$ then for every $x, h \in \mathbb{R}^d$

$$f(x+h) \;=\; f(x) \;+\; \langle \nabla f(\tilde{x}), h \rangle$$

where $\tilde{x} = x + \alpha h$ for some $\alpha \in [0, 1]$. In particular, we have

$$f(x+h) \;=\; f(x) \;+\; \langle \nabla f(\tilde{x}), h \rangle \;+\; o(||h||)$$

Multivariate Taylor's Theorem II

Fact: If $f : \mathbb{R}^d \to \mathbb{R}$ has continuous second partial derivatives $\partial^2 f / \partial x_i \partial x_j$ at each point in $\mathbb{R}^d$ then for every $x, h \in \mathbb{R}^d$

$$f(x+h) \;=\; f(x) \;+\; \langle \nabla f(x), h \rangle \;+\; \frac{1}{2} h^t \nabla^2 f(\tilde{x}) h$$

where $\tilde{x} = x + \alpha h$ for some $\alpha \in [0, 1]$. In particular, we have

$$f(x+h) \;=\; f(x) \;+\; \langle \nabla f(x), h \rangle \;+\; \frac{1}{2} h^t \nabla^2 f(x) h \;+\; o(||h||^2)$$