Core Concepts and Theorems in Classical Optimization Theory

(Complete Compilation of Part I: Optimization Theory)

Compiled by Gemini AI Assistant

November 2025

Contents

1 Unconstrained Optimization Concepts and Theorems

1.1 Unconstrained Optimization Concepts (Definitions)

1. **Unconstrained Optimization Problem** The problem of finding $\min_{x\in\mathbb{R}^n} f(x)$ or $f(x) \to \min_{x\in\mathbb{R}^n}$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth (continuously differentiable) scalar function.

2. **Global Minimum** A point $x^* \in \mathbb{R}^n$ is a global minimum of f if and only if $f(x^*) \leq f(x)$ holds for all $x \in \mathbb{R}^n$.

3. **Local Minimum** A point $x^* \in \mathbb{R}^n$ is a local minimum of f if and only if there exists $\epsilon > 0$ such that $f(x^*) \leq f(x)$ holds for all x satisfying $||x - x^*|| < \epsilon$.

4. **Strict Local Minimum** A vector $x^* \in \mathbb{R}^n$ is a strict local solution if there exists a neighborhood $\mathcal{N}$ such that $f(x) > f(x^*)$ holds for all x satisfying $x \in \mathcal{N} \cap \Omega, x \neq x^*$.

5. **Stationary Point** A point $x^* \in \mathbb{R}^n$ is a stationary point if and only if $\nabla f(x^*) = 0$.

6. **Descent Direction** For a function f and a point x, a vector $p \in \mathbb{R}^n$ is called a descent direction if it satisfies $\nabla f(x)^T p < 0$.

7. **Gradient (∇f)** ∇f is the gradient vector of the function f.

8. **Hessian ($\nabla^2 f$)** $\nabla^2 f$ is the Hessian matrix of the function f.

9. **Positive Definiteness** A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if and only if $p^T Ap > 0$ for all $p \neq 0$.

10. **Positive Semidefiniteness** A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if $p^T Ap \geq 0$ for all $p \in \mathbb{R}^n$.

1.2 Relevant Theorems

Theorem 1.1 (Taylor's Theorem). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, and $p \in \mathbb{R}^n$. Then there exists $t \in (0, 1)$ such that:*

$$f(x + p) = f(x) + \nabla f(x + tp)^T p$$

If f is twice continuously differentiable, then there exists $t \in (0, 1)$ such that:

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp)p$$

Theorem 1.2 (Necessary Condition for a Local Minimum). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open neighborhood of x^*. If x^* is a local minimum of f, then:*

$$\nabla f(x^*) = 0$$

Theorem 1.3 (Second-Order Necessary Condition). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in an open neighborhood of x^*. If x^* is a local minimum of f, then:*

$$\nabla f(x^*) = 0 \quad \textit{and} \quad \nabla^2 f(x^*) \textit{ is positive semidefinite.}$$

Theorem 1.4 (Second-Order Sufficient Condition). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in an open neighborhood of x^*. Suppose:*

$$\nabla f(x^*) = 0 \quad \textit{and} \quad \nabla^2 f(x^*) \textit{ is positive definite.}$$

Then x^* is a strict local minimum of f.

Theorem 1.5 (Global Minimum for Convex Functions). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then any local minimizer x^* is also a global minimizer. If in addition f is differentiable, then any stationary point x^* with $\nabla f(x^*) = 0$ is a global minimizer.*

2 Line-Search Concepts and Theorems

2.1 Line-Search Concepts (Definitions)

1. **Descent Direction** For a function f and a point x, a vector $p \in \mathbb{R}^n$ is called a descent direction if it satisfies $\nabla f(x)^T p < 0$.

2. **Step Length (α)** In the iteration $x_{k+1} = x_k + \alpha_k p_k$, $\alpha_k > 0$ is the distance moved along the search direction p_k.

3. **Sufficient Decrease Condition (Armijo Condition)** A step length $\alpha > 0$ satisfies the sufficient decrease condition if and only if:
$$f(x + \alpha p) \leq f(x) + c_1 \alpha \nabla f(x)^T p$$
where $c_1 \in (0, 1)$.

4. **Curvature Condition** A step length $\alpha > 0$ satisfies the curvature condition if and only if:
$$\nabla f(x + \alpha p)^T p \geq c_2 \nabla f(x)^T p$$
where $c_2 \in (c_1, 1)$.

5. **Wolfe Conditions** A step length $\alpha > 0$ satisfies the Wolfe conditions if and only if it satisfies both:
   (a) Sufficient decrease: $f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T p_k$
   (b) Curvature condition: $\nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f(x_k)^T p_k$

   where $0 < c_1 < c_2 < 1$.

6. **Strong Wolfe Conditions** A step length $\alpha > 0$ satisfies the strong Wolfe conditions if and only if it satisfies the sufficient decrease condition (1) and replaces the curvature condition (2) with:
$$|\nabla f(x_k + \alpha p_k)^T p_k| \leq c_2 |\nabla f(x_k)^T p_k|$$
where $0 < c_1 < c_2 < 1$.

7. **Goldstein Conditions** A step length $\alpha > 0$ satisfies the Goldstein conditions if and only if:
$$f(x_k) + (1 - c)\alpha \nabla f(x_k)^T p_k \leq f(x_k + \alpha p_k) \leq f(x_k) + c\alpha \nabla f(x_k)^T p_k$$
where $c \in (0, 0.5)$.

2.2 Relevant Results

Lemma 2.1 (Existence of Step Lengths). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Suppose p_k is a descent direction at x_k, and f is bounded below along the ray $\{x_k + \alpha p_k | \alpha > 0\}$. Then, for any $0 < c_1 < c_2 < 1$, there exists an interval of step lengths α such that the Wolfe and strong Wolfe conditions are satisfied.*

Theorem 2.2 (Global Convergence of Line Search Methods). *Let $\{x_k\}$ be a sequence generated by the iteration $x_{k+1} = x_k + \alpha_k p_k$ where p_k is a descent direction, and the step length α_k satisfies Wolfe conditions. Assume:*

1. *The function f is bounded below on $\mathbb{R}^n$.*

2. *f is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\mathcal{L} = \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}$, where x_0 is the starting point of the iteration.*

3. *The gradient ∇f is Lipschitz continuous on $\mathcal{N}$, i.e., $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for all $x, y \in \mathcal{N}$.*

Then the Zoutendijk condition holds:
$$\sum_{k \geq 0} (\cos \theta_k)^2 ||\nabla f(x_k)||^2 < \infty$$
where $\cos \theta_k = \frac{-\nabla f(x_k)^T p_k}{||\nabla f(x_k)|| ||p_k||}$.

3 Newton and Quasi-Newton Methods (Statements)

3.1 Core Concepts

1. **Newton Step Definition** The search direction for Newton's method is defined as $p_k^N = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$.

2. **Local Quadratic Convergence Statement** If the initial point x_0 is sufficiently close to x^* (where $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$), the sequence of iterates $\{x_k\}$ generated by the pure Newton method converges to x^* at a Q-quadratic rate, i.e., $||x_{k+1} - x^*|| \leq M||x_k - x^*||^2$ for some constant M.

3. **The Hessian Modification Idea** If the Hessian $\nabla^2 f(x_k)$ is not positive definite, it is replaced by a modified matrix $B_k = \nabla^2 f(x_k) + E_k$, where E_k is chosen such that B_k is positive definite, ensuring that $p_k = -B_k^{-1}\nabla f(x_k)$ is a descent direction.

4. **Quasi-Newton Secant Condition** The updated Quasi-Newton matrix B_{k+1} must satisfy the secant equation: $B_{k+1}s_k = y_k$, where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$.

3.2 Relevant Theorems/Lemmas

Theorem 3.1 (Local Convergence of Newton's Method)**.** *Suppose f is twice differentiable and that $\nabla^2 f(x)$ is Lipschitz continuous near a solution x^* satisfying the second-order sufficient conditions. Consider the iteration $x_{k+1} = x_k + p_k$ where p_k is the Newton step. Then:*

1. *If the starting point x_0 is sufficiently close to x^*, the sequence of iterates converges to x^*.*

2. *The convergence rate of $\{x_k\}$ is quadratic.*

3. *The sequence $\{||\nabla f_k||\}$ converges quadratically to zero.*

Theorem 3.2 (Quasi-Newton / Superlinear Convergence Statements)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Let the iteration $x_{k+1} = x_k + \alpha_k p_k$ be generated by a descent direction p_k and a step length α_k satisfying Wolfe conditions (with $c_1 \leq \frac{1}{2}$). If the sequence $\{x_k\}$ converges to x^* such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, and if the search direction satisfies:*

$$\lim_{k\to\infty} \frac{||\nabla f(x_k) + \nabla^2 f(x_k)p_k||}{||p_k||} = 0$$

then:

1. *The step length $\alpha_k = 1$ is admissible for all k greater than a certain index k_0.*

2. *If $\alpha_k = 1$ for all $k > k_0$, then $\{x_k\}$ converges to x^* superlinearly.*

4 Trust-Region Methods

4.1 Core Concepts

1. **The Quadratic Trust-Region Model** The model is $m_k(p) = f_k + g_k^T p + \frac{1}{2}p^T B_k p$ minimized within the constraint $||p|| \leq \Delta_k$.

2. **Trust-Region Radius Strategy** The strategy chooses the trust-region radius Δ_k based on the agreement between the model m_k and the objective f, using the ratio $\rho_k = \frac{f(x_k)-f(x_k+p_k)}{m_k(0)-m_k(p_k)}$.

3. **The Cauchy Point (Closed-Form)** The Cauchy point p_k^c is defined as $-\tau_k \frac{\Delta_k}{||g_k||} g_k$, where:

$$\tau_k = \begin{cases} 1, & \text{if } g_k^T B_k g_k \leq 0; \\ \min(\frac{||g_k||^3}{\Delta_k g_k^T B_k g_k}, 1), & \text{otherwise.} \end{cases}$$

4. **Dogleg Path Idea** The Dogleg Method constructs a piecewise linear path from the origin to the unconstrained minimum along the steepest descent direction (p^U), and then toward the full Newton step (p^B), to approximate the trust-region solution.

4.2 Relevant Results

Lemma 4.1 (Existence of Global Minimum of Unconstrained Quadratic Model)**.** *Let $m(p) = g^T p + \frac{1}{2}p^T Bp$, where B is symmetric.*

1. *m attains a minimum if and only if B is positive semidefinite and g is in the range of B. If B is positive semidefinite, then every p satisfying $Bp = -g$ is a global minimizer of m.*

2. *m has a unique minimizer if and only if B is positive definite.*

Lemma 4.2 (Dogleg Properties)**.** *Let B be positive definite. Then along the dogleg path $\tilde{p}(\tau)$:*

1. *$||\tilde{p}(\tau)||$ is an increasing function of τ.*

2. *$m(\tilde{p}(\tau))$ is a decreasing function of τ.*

Lemma 4.3 (Cauchy Decrease)**.** *The Cauchy point p_k^c satisfies the bound:*

$$m_k(0) - m_k(p_k^c) \geq \frac{1}{2}||g_k|| \min\left(\Delta_k, \frac{||g_k||}{||B_k||}\right)$$

Theorem 4.4 (Characterization of Trust-Region Solution / Moré-Sorensen Type Result)**.** *The vector p^* is a global solution of the trust-region subproblem* $\min_{p\in\mathbb{R}^n} m(p)$ *s.t.* $||p|| \leq \Delta$*, if and only if p^* is feasible and there exists a scalar $\lambda \geq 0$ such that:*

1. $(B + \lambda I)p^* = -g$

2. $\lambda(\Delta - ||p^*||) = 0$

3. *$B + \lambda I$ is positive semidefinite*

5 Conjugate Gradient and Conjugate Directions

5.1 Core Concepts

1. **Definition of A-conjugacy** A set of nonzero vectors $\{p_0, p_1, \ldots, p_l\}$ is conjugate with respect to an $n \times n$ symmetric positive definite matrix A if it satisfies $p_i^T A p_j = 0$ for all $i \neq j$.

2. **Krylov Subspaces** A Krylov subspace $\mathcal{K}(r_0; k)$ is defined as span$\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$.

3. **Finite Termination Property** If the matrix A has only r distinct eigenvalues, the Conjugate Gradient (CG) iteration will terminate at the solution in at most r iterations.

4. **Orthogonality of Residuals** The sequence of residuals $\{r_k\}$ generated by the CG method satisfies $r_k^T r_i = 0$ for $i \neq k$, i.e., the residuals are mutually orthogonal.

5. **Spectral Convergence Bounds** The convergence rate bounds depend on the distribution of eigenvalues of the matrix A.

6. **Preconditioning Concept** Preconditioning involves transforming the original system $Ax = b$ into an equivalent system with a better-conditioned matrix (e.g., $M^{-1}A$ or $C^{-T}AC^{-1}$) to accelerate the convergence rate of CG.

5.2 Relevant Theorems

Theorem 5.1 (Convergence of Conjugate Direction Method)**.** *For any $x_0 \in \mathbb{R}^n$, the sequence $\{x_k\}$ generated by the conjugate direction algorithm converges to the solution x^* of the linear system $Ax = b$ in at most n steps.*

Theorem 5.2 (Finite-Termination Property)**.** *If A has only r distinct eigenvalues, then the Conjugate Gradient (CG) iteration will terminate at the solution in at most r iterations.*

6 Constrained Optimization

6.1 Core Definitions

1. **Problem Statement** Minimize $f(x)$ over $x \in \mathbb{R}^n$ subject to $c_i(x) = 0, i \in E$ (equality constraints) and $c_i(x) \geq 0, i \in I$ (inequality constraints).

2. **Feasible Set (Ω)** The feasible set Ω is the set of all vectors $x \in \mathbb{R}^n$ satisfying all constraints.

3. **Feasible Point** A feasible point is any vector $x \in \mathbb{R}^n$ belonging to the feasible set Ω.

4. **Active Set ($A(x)$)** The active set $A(x)$ at a feasible point x consists of the equality constraint indices E together with the indices of the inequality constraints $i \in I$ for which $c_i(x) = 0$.

5. **Active Constraint** An inequality constraint $i \in I$ is active at x if $c_i(x) = 0$. All equality constraints $i \in E$ are always active.

6. **Inactive Constraint** An inequality constraint $i \in I$ is inactive at x if the strict inequality $c_i(x) > 0$ is satisfied.

7. **Feasible Sequence** A sequence $\{z_k\}$ approaching x is called a feasible sequence if $z_k \in \Omega$ for all k sufficiently large and $z_k \to x$.

8. **Tangent Cone ($T_\Omega(x)$)** The set of all limiting directions d of feasible sequences approaching x, where $d = \lim_{k\to\infty} \frac{z_k - x}{t_k}$ for some feasible sequence $\{z_k\}$ and positive scalars $\{t_k\}$ with $t_k \to 0$.

9. **Set of Linearized Feasible Directions ($F(x)$)** The set of vectors d satisfying:

$$F(x) = \left\{ d \mid d^T \nabla c_i(x) = 0, \text{ for all } i \in E; \quad d^T \nabla c_i(x) \geq 0, \text{ for all } i \in A(x) \cap I \right\}$$

10. **LICQ (Linear Independence Constraint Qualification)** LICQ holds at point x with active set $A(x)$ if the set of active constraint gradients $\{\nabla c_i(x), i \in A(x)\}$ is linearly independent.

11. **MFCQ (Mangasarian-Fromovitz Constraint Qualification)** MFCQ holds, if there exists $w \in \mathbb{R}^n$ such that $\nabla c_i(x^*)^T w > 0$ for all $i \in A(x^*) \cap I$, $\nabla c_i(x^*)^T w = 0$ for all $i \in E$, and the set $\{\nabla c_i(x^*), i \in E\}$ is linearly independent.

7 Lagrangian, KKT, and First-Order Optimality

7.1 Core Concepts

1. **The Lagrangian Function** The Lagrangian is defined as $\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in E \cup I} \lambda_i c_i(x)$.

2. **KKT Conditions** The necessary conditions for a local solution x^* (assuming LICQ) require the existence of λ^* satisfying:

   (a) **Stationarity**: $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$.

   (b) **Primal Feasibility**: $c_i(x^*) = 0, i \in E$ and $c_i(x^*) \geq 0, i \in I$.

   (c) **Dual Feasibility**: $\lambda_i^* \geq 0$ for all $i \in I$.

   (d) **Complementary Slackness**: $\lambda_i^* c_i(x^*) = 0$ for all $i \in E \cup I$.

3. **Role of Constraint Qualifications for Necessity/Sufficiency** Constraint Qualifications guarantee that the KKT conditions are **necessary** for a local solution.

7.2 Relevant Items in the Notes

Lemma 7.1 (Tangent Cone and First-Order Feasible Directions)**.** *$T_\Omega(x^*) \subset F(x^*)$, and if LICQ holds at x^*, then $T_\Omega(x^*) = F(x^*)$.*

Theorem 7.2 (First-Order Necessary Conditions)**.** *Suppose x^* is a local solution, f and c_i are continuously differentiable, and that LICQ holds at x^*. Then there is a Lagrange multiplier vector λ^* satisfying the KKT conditions.*

8 Second-Order Optimality for Constrained Problems

8.1 Core Concepts

1. **Critical Cone Definition** The critical cone $C(x^*, \lambda^*)$ contains directions $w \in F(x^*)$ for which $w^T \nabla f(x^*) = 0$, defined by the conditions in Section 7.2.1.

2. **Second-Order Necessary Conditions** They require $w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \geq 0$ for all $w \in C(x^*, \lambda^*)$.

3. **Second-Order Sufficient Conditions** They require $w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w > 0$ for all $w \in C(x^*, \lambda^*), w \neq 0$.

4. **Normal Cone Definition** The normal cone $N_\Omega(x)$ is defined as $N_\Omega(x) = \{v \mid v^T w \leq 0 \text{ for all } w \in T_\Omega(x)\}$.

8.2 Relevant Theorems

Theorem 8.1 (Second-Order Necessary Conditions)**.** *Suppose x^* is a local solution and that LICQ holds. Let λ^* be a Lagrange multiplier satisfying the KKT conditions. Then $w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w \geq 0$ for all $w \in C(x^*, \lambda^*)$.*

Theorem 8.2 (Second-Order Sufficient Conditions)**.** *Suppose that for some feasible point x^*, there exists a multiplier λ^* such that the KKT conditions are satisfied. Further suppose that $w^T \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) w > 0$ for all $w \in C(x^*, \lambda^*), w \neq 0$. Then x^* is a strict local solution of the problem.*

9 Duality Theory (Nonlinear and LP)

9.1 Core Concepts

1. **Dual Function** $q(\lambda) = \inf_x[f(x) - \lambda^T c(x)]$.
2. **Weak Duality** Weak duality states that $q(\bar{\lambda}) \le f(\bar{x})$ for any primal feasible $\bar{x}$ and dual feasible $\bar{\lambda} \ge 0$.
3. **Duality for Convex Quadratic Programming** The dual problem maximizes a concave quadratic objective subject to nonnegativity constraints.
4. **Wolfe Dual (Construction and Interpretation)** The Wolfe dual is $\max_{x,\lambda} \mathcal{L}(x, \lambda)$ s.t. $\nabla_x \mathcal{L}(x, \lambda) = 0$ and $\lambda \ge 0$.

9.2 Relevant Theorems

Theorem 9.1 (Concavity of the Dual Objective)**.** *The dual objective $q(\lambda) = \inf_x \mathcal{L}(x, \lambda)$ is concave, and its domain $D = \{\lambda \mid q(\lambda) > -\infty\}$ is convex.*

Theorem 9.2 (Weak Duality)**.** *For any feasible $\bar{x}$ in $\min_x f(x)$ s.t. $c(x) \ge 0$ and $\bar{\lambda} \ge 0$, $q(\bar{\lambda}) \le f(\bar{x})$.*

Theorem 9.3 (Solutions of the Dual Problem)**.** *If $\bar{x}$ solves $\min_x f(x)$ s.t. $c(x) \ge 0$, and $f, -c_i$ are convex and differentiable at $\bar{x}$, then any $\bar{\lambda}$ satisfying the KKT conditions with $\bar{x}$ is a solution of the dual problem.*

Theorem 9.4 (Wolfe Dual Formulation)**.** *If f and $-c_i$ are convex and continuously differentiable, and $(\bar{x}, \bar{\lambda})$ satisfies the KKT conditions, then $(\bar{x}, \bar{\lambda})$ solves the Wolfe Dual $\max_{x,\lambda} \mathcal{L}(x, \lambda)$ s.t. $\nabla_x \mathcal{L}(x, \lambda) = 0, \lambda \ge 0$. </theorem*

10 Linear Programming and the Simplex Method (Concepts)

10.1 Core Concepts

1. ***Standard Form of LP***
$$\min c^T x, \quad \textit{s.t.} \quad Ax = b, \quad x \geq 0$$

2. ***Vertices and Basic Feasible Points*** *All basic feasible points for the LP problem are vertices of the feasible polytope* $\{x \mid Ax = b, x \geq 0\}$*, and vice versa.*

3. ***Statement of the Fundamental Theorem of Linear Programming*** *See Theorem 10.1.*

10.2 Relevant Results

Theorem 10.1 (Fundamental Theorem of Linear Programming)**.**

1. *If the primal problem has a nonempty feasible region, then there is at least one basic feasible point.*

2. *If the primal problem has solutions, then at least one solution is a basic optimal point.*

3. *If the nonempty feasible region is bounded, then the primal problem has an optimal solution.*

11 Quadratic Programming (QP)

11.1 Core Concepts

1. ***Definition of a Quadratic Program*** *A QP minimizes a quadratic objective function* $q(x) = \frac{1}{2}x^T Gx + x^T c$ *subject to linear constraints.*

2. ***Equality-Constrained QP*** *A QP where all constraints are linear equalities* $Ax = b$*.*

3. ***KKT system for Equality-Constrained QP*** *The KKT system is:*
$$\begin{pmatrix} G & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}$$

4. ***Sufficiency Conditions for Convex QP*** *For a convex QP (G is positive semidefinite), KKT conditions are sufficient for global optimality.*

11.2 Relevant Theorems

Theorem 11.1 (Global Solution for the Equality-Constrained QP)**.** *If A has full row rank and* $Z^T GZ$ *is positive definite (where* Z *is the null-space basis for A), then* x^* *satisfying the KKT system is the unique global solution of the equality-constrained QP.*

Theorem 11.2 (Sufficiency for Convex QP)**.** *If* x^* *satisfies KKT conditions with* λ_i^**, and G is positive semidefinite (convex QP), then* x^* *is a global solution of the QP.*

12 Descriptive Regression and Least Squares Estimation (LSE)

12.1 Descriptive Regression General Setting

1. ***Regression Purpose*** *To represent and understand the relationship between a response (output) and several influencing factors (inputs). Regression analysis is a statistical method for estimating relationships among variables, describing how a dependent variable changes as one or more independent variables change.*

2. ***Dependent Variable (y)*** *The outcome or response variable we are trying to predict or explain.*

3. ***Independent Variables ($x_1, \ldots, x_m$)*** *The predictor variables or factors that influence the dependent variable. Also called regressors.*

4. ***Model Function ($\eta(x_1, \ldots, x_m)$)*** *A mathematical representation of the relationship: $y = \eta(x_1, \ldots, x_m)$.*

5. ***Role of Regression Analysis as a Descriptive Modeling Tool*** *The focus of descriptive regression is to build a model based on observed data without making strong assumptions about the underlying statistical distribution.*

12.2 Parametric Regression and Least Squares Formulation

1. ***Idea of Model Fitting*** *The core challenge is that exact functional relationships are often too complex to determine, so we use simplified parametric models $\eta(x, \theta)$ that approximate the statistical dependence with satisfactory accuracy.*

2. ***Residuals and Error Minimization*** *The deviation (error) for the j-th observation is $\epsilon_j = y_j - \tilde{y}_j$, where $\tilde{y}_j$ are the predicted values. We seek to minimize the magnitude of the error vector $\epsilon = Y - X\theta$.*

3. ***Least Squares Criterion and its Justification*** *The criterion is to minimize the sum of squared errors: $\sum \epsilon_j^2 = \epsilon^T \epsilon = (Y - X\theta)^T (Y - X\theta)$. This is preferred due to its computational simplicity and strong optimality properties (e.g., resulting in a convex optimization problem).*

12.3 Least Squares Estimation (LSE)

1. ***Definition of the Least Squares Estimator (LSE)*** *The vector $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^m} (Y - X\theta)^T (Y - X\theta) = \arg\min_{\theta \in \mathbb{R}^m} \epsilon^T \epsilon$ is called the empirical least squares estimator.*

2. ***System of Normal Equations*** *The system of equations $X^T X\theta = X^T Y$ provides a closed-form condition for the optimality of θ.*

3. ***Convexity of the Least Squares Criterion*** *The least squares criterion $f(\theta) = (Y - X\theta)^T (Y - X\theta)$ is quadratic in θ, leading to a convex optimization problem.*

Lemma 12.1 (Normal Equations Existence and Optimality Condition)**.** *For any matrix X and vector Y of compatible dimensions, the system of normal equations $X^T X\theta = X^T Y$ always has at least one solution. Any vector $\hat{\theta}$ satisfying this system is a least squares estimator.*

13 Classical Linear Regression Model and BLUE

13.1 Classical Linear Regression Model

1. ***Definition of the Classical Linear Regression Model*** *The model is written as* $y_j = \sum_{i=1}^{m} x_{ji}\theta_i + \epsilon_j$ *for* $j = 1, ..., N$*, or in matrix form:* $Y = X\theta + \epsilon$.

2. ***Assumptions of Zero Mean, Uncorrelated and Homoscedastic Errors*** *The error vector* ϵ *satisfies:*

   - *Zero Mean (Unbiasedness):* $E[\epsilon_i] = 0$.
   - *Homoscedasticity:* $E[\epsilon_i^2] = \sigma^2$ *(common variance).*
   - *Uncorrelated Errors:* $E[\epsilon_i\epsilon_j] = 0$ *for* $i \neq j$.

   The covariance matrix is $Cov(\epsilon) = \Sigma = E[\epsilon\epsilon^T] = \sigma^2 I_N$.

3. ***Definition of OLS (Ordinary Least Squares) Estimator*** *For a classical linear regression model with a non-singular matrix* X^TX*, the vector* $\hat{\theta} = (X^TX)^{-1}X^TY$ *is called the Ordinary Least Squares estimator.*

13.2 Best Linear Unbiased Estimator (BLUE)

1. ***Definition and Conditions for Linear Unbiased Estimators*** *A linear estimator* $\tilde{\theta} = AY$ *is unbiased if and only if* $AX = I$.

2. ***Covariance Dominance (Minimum Variance)*** *For any vector* z *of appropriate dimension and any unbiased estimator* $\tilde{\theta}$*, the inequality* $D(z^T(\hat{\theta} - \theta)) \leq D(z^T(\tilde{\theta} - \theta))$ *holds, where* D *denotes the covariance matrix, and* $\hat{\theta}$ *is the BLUE. This implies that the matrix* $D_{\tilde{\theta}} - D_{\hat{\theta}}$ *is positive semi-definite.*

3. ***Conditions (a)-(c) for BLUE*** *An estimator* $\hat{\theta}$ *is the Best Linear Unbiased Estimator (BLUE) if it satisfies:*

   - *(a) Estimator Unbiasedness:* $E[\hat{\theta}] = \theta$.
   - *(b) Minimum Variance:* $\hat{\theta}$ *minimizes the variance of any linear combination* $z^T\hat{\theta}$ *among all linear unbiased estimators.*
   - *(c) Linearity:* $\hat{\theta} = SY$*, where* S *is a matrix independent of* Y.

4. ***Variance Minimization Principle*** *The BLUE minimizes the variance of every linear combination* $z^T\hat{\theta}$ *across all possible linear unbiased estimators.*

Theorem 13.1 (Gauss-Markov Theorem)**.** *Consider the classical linear regression model* $(Y, X\theta, \sigma^2 I_N)$*, where the error vector* ϵ *satisfies: (1)* $E[\epsilon] = 0$*, (2)* $Cov(\epsilon) = \sigma^2 I_N$*, and (3) components are uncorrelated. If the matrix* X^TX *is nonsingular, then the vector*

$$\hat{\theta} = (X^TX)^{-1}X^TY$$

is the Best Linear Unbiased Estimator (BLUE) of θ*. Its covariance matrix is given by* $D_{\hat{\theta}} = \sigma^2(X^TX)^{-1}$.

Lemma 13.2 (Properties of the OLS Estimator)**.** *The OLS estimator* $\hat{\theta} = (X^TX)^{-1}X^TY$ *is a linear and unbiased estimator, i.e., it satisfies conditions (a) and (c) for BLUE.*

Lemma 13.3 (Unbiasedness Condition)**.** *A linear estimator* $\tilde{\theta} = AY$ *is unbiased if and only if* $AX = I$.

Lemma 13.4 (Covariance Form)**.** *Under the assumptions of the Gauss-Markov Theorem, the covariance matrix of any linear unbiased estimator* $\tilde{\theta} = AY$ *is* $D_{\tilde{\theta}} = \sigma^2 AA^T$.

14 OLS Estimation in the Singular Case

14.1 OLS Estimation in the Singular Case

1. ***OLS solutions for singular*** X^TX *If* X^TX *is singular (rank*$(X^TX) < m$*), the normal equation* $X^TX\theta = X^TY$ *still has solutions, but not a unique one. Any vector* $\hat{\theta}$ *satisfying this equation minimizes the residual sum of squares* $||Y - X\theta||^2$ *and is called an OLS estimator.*

2. ***Concept of the Generalized Inverse*** *A matrix* $A^- \in \mathbb{R}^{m\times n}$ *is called a generalized inverse of* A *if, for every vector* $y \in \mathbb{R}^n$ *such that the system* $Ax = y$ *is consistent, the vector* $x = A^-y$ *is a solution.*

3. ***Moore-Penrose Pseudoinverse*** *A matrix* $A^+ \in \mathbb{R}^{m\times n}$ *is called the Moore-Penrose pseudoinverse of* A *if it satisfies the four Penrose conditions: (1)* $AA^+A = A$*, (2)* $A^+AA^+ = A^+$*, (3)* $(AA^+)^\top = AA^+$*, and (4)* $(A^+A)^\top = A^+A$*.*

4. ***Penrose Conditions*** *The four defining algebraic conditions for the Moore-Penrose pseudoinverse (listed above).*

5. ***Representation of all OLS estimators via the Generalized Inverse*** *The general solution to the consistent system* $X^TX\theta = X^TY$ *can be represented using a generalized inverse* $A^- = (X^TX)^-$ *as* $\theta = A^-X^TY + (H - I)z$*, where* $H = A^-A$ *and* z *is an arbitrary vector.*

Theorem 14.1 (Generalized Inverse Solution)**.** *A linear parametric function* $\tau = T\theta$ *is estimable if and only if* $T(X^TX)^-X^TX = T$*. If this condition is satisfied, the OLS-estimator* $\hat{\tau} = T(X^TX)^-X^TY$ *is uniquely defined and represents the best linear unbiased estimator.*

Lemma 14.2 (Condition for Generalized Inverse)**.** *For a matrix* B *to be a generalized inverse of matrix* A*, it is necessary and sufficient that* $ABA = A$*.*

Lemma 14.3 (Existence of Generalized Inverse)**.** *For any matrix* A*, there exists a generalized inverse* A^-*.*

15 WLS/GLS and Design of Experiments

15.1 Weighted and Generalized Least Squares (WLS / GLS)

1. ***Generalized Linear Regression Model*** *The model is written as* $(Y, X\theta, \sigma^2 W)$*, where* $W \in \mathbb{R}^{N\times N}$ *is a known positive definite matrix, and* $\sigma^2 > 0$ *is an unknown scalar parameter.*

2. ***Definition of the GLS Estimator*** *The Generalized Least Squares (GLS) estimator is defined as:*

$$\hat{\theta} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$$

3. ***Properties of the GLS Estimator*** *The covariance matrix of* $\hat{\theta}$ *is* $D_{\hat{\theta}} = \sigma^2 (X^T W^{-1} X)^{-1}$.

4. ***Optimality (BLUE in the Generalized Model)*** *According to the Gauss-Markov theorem, the GLS estimator is the Best Linear Unbiased Estimator (BLUE) under the generalized linear model assumptions.*

15.2 Design of Experiments: Basic Concepts

1. ***Model Structure*** $y_j = \eta(t_j, \theta) + \epsilon_j$ *for* $j = 1, ..., N$*, where* $\eta(t, \theta) = \theta^T f(t)$ *(linearity in* θ*),* $t_j \in \chi$ *(design points), and* ϵ_j *are observation errors.*

2. ***Standard Assumptions (a)-(f)*** *The model assumes (a) Unbiasedness (*$E[\epsilon_j] = 0$*), (b) Uncorrelated errors (*$E[\epsilon_i \epsilon_j] = 0$ *for* $i \neq j$*), (c) Homoscedasticity (*$E[\epsilon_j^2] = \sigma^2 > 0$*), (d) Linearity in* θ*, (e) Basis functions* $f_i(t)$ *are continuous and linearly independent on* χ*, and (f) Design space* χ *is compact.*

3. ***Definition of Discrete Design (***ξ_N***)*** *A design represented by* $\xi_N = \left(\begin{smallmatrix} t_1 & \cdots & t_N \\ 1/N & \cdots & 1/N \end{smallmatrix}\right)$*, where* t_i *may repeat, and* N *is the total number of observations.*

4. ***Definition of Approximate Design (***ξ***)*** *A design represented by* $\xi = \left(\begin{smallmatrix} t_1 & \cdots & t_n \\ \omega_1 & \cdots & \omega_n \end{smallmatrix}\right)$*, where* t_i *are* n *distinct support points,* $\omega_i \geq 0$ *are weights (relative frequencies), and* $\sum_{i=1}^n \omega_i = 1$.

5. ***Full Design Space (***Ξ***)*** *The full design space* Ξ *is the union of all approximate designs with exactly* n *support points:* $\Xi = \bigcup_{n=1}^{\infty} \Xi_n$.

15.3 Information Matrix and Variance of Estimates

1. ***Definition of the Information Matrix (***$M(\xi)$***)*** *The information matrix of a design* ξ *is defined as the integral of the outer product of the regressor vector* $f(t)$ *with respect to the design measure* $\xi(dt)$*:*

$$M(\xi) = \int_{\chi} f(t) f^T(t) \xi(dt) \in \mathbb{R}^{m\times m}$$

2. ***Relation to Covariance of Parameter Estimates*** *For a nonsingular design* ξ*, the information matrix is proportional to the inverse of the covariance matrix of the OLS estimator:* $D_{\hat{\theta}} = \sigma^2 N^{-1} (M(\xi))^{-1}$.

3. ***Criteria of Optimality based on*** $M(\xi)$ *Optimality criteria are real-valued functions defined on* $M(\xi)$ *(or its inverse* $D(\xi) = M(\xi)^{-1}$*) that quantify the precision of the design.*

Theorem 15.1 (Properties of Information Matrices)**.** *The following statements hold for the set* $\mathcal{M} = \{M : M = M(\xi), \xi \in \Xi\}$*:*

1. *Every information matrix* $M(\xi)$ *is positive semidefinite.*

2. *If* $n < m$ *(fewer support points than parameters), then* $\det M(\xi) = 0$.

3. *The set* $\mathcal{M}$ *of all information matrices is convex.*

4. *Under regularity conditions,* $\mathcal{M}$ *is compact.*

5. *For any* $\xi \in \Xi$*, there exists* $\tilde{\xi}$ *with* $n \leq \frac{m(m+1)}{2} + 1$ *support points such that* $M(\tilde{\xi}) = M(\xi)$.

16 Optimality Criteria and Equivalence Theorem

16.1 Optimality Criteria in Design

1. ***D-optimality Criterion*** *Minimizes the volume of the confidence ellipsoid by maximizing the determinant of the information matrix:* $\log\det M(\xi) \to \sup_{\xi}$ *or* $\log\det D(\xi) \to \inf_{\xi\in\Xi}$.

2. ***L-optimality Criterion*** *Minimizes generalized quadratic loss by minimizing the trace of* L *times the inverse information matrix: tr*$[LD(\xi)] \to \inf_{\xi\in\Xi_{NS}}$*, where* L *is a fixed nonnegative definite matrix.*

3. ***E-optimality Criterion*** *Maximizes the smallest eigenvalue of the information matrix, which minimizes the longest axis of the confidence ellipsoid:* $\lambda_{\min}(M(\xi)) \to \sup_{\xi\in\Xi_{NS}}$ *or* $\lambda_{\max}(D(\xi)) \to \inf_{\xi\in\Xi_{NS}}$.

4. e_k***-optimality Criterion*** *Minimizes the variance of the estimator for a single parameter coordinate* θ_k*:* $e_k^T M^-(\xi) e_k \to \inf_{\xi\in\Xi_{e_k}}$.

5. ***G-optimality Criterion*** *Minimizes the maximum prediction variance over the design space:* $\max_{t\in\chi} d(t,\xi) \to \inf_{\xi\in\Xi_{NS}}$*, where* $d(t,\xi) = f^T(t)M^{-1}(\xi)f(t)$.

Theorem 16.1 (Kiefer-Wolfowitz Equivalence Theorem)**.** *Under standard regularity assumptions (including compactness of the set of information matrices), the following conditions are equivalent for a design* ξ^**:*

1. ξ^* *is D-optimal.*

2. ξ^* *is G-optimal.*

3. $\max_{x\in\chi} d(x,\xi^*) = m$*, where* m *is the number of parameters.*

Theorem 16.2 (D-Optimal Designs for the Polynomial Model)**.** *For polynomial regression of degree* $m-1$ *on* $\chi = [-1,1]$*, a unique approximate D-optimal design exists. It is supported with equal weights (*$1/m$*) on* m *points, which are the roots of the polynomial* $(x^2-1)P'_{m-1}(x)$*, where* $P_{m-1}(x)$ *is the Legendre polynomial of degree* $m-1$.

Theorem 16.3 (D-Optimal Designs for the Trigonometric Model)**.** *An approximate D-optimal design for the trigonometric regression model of order* m *on* $\chi = [-\pi,\pi]$ *is any design supported with equal weights (*$1/N$*) on* $N \geq 2m+1$ *equally spaced points* $t_i^* = \frac{i-1}{N}2\pi - \pi$ *for* $i = 1,\ldots,N$*. The continuous uniform design* $\xi^* = \frac{1}{2\pi}dx$ *is also D-optimal.*