DSA3102 Cheatsheet

Ng Xuan Jun

Unconstrained Problem Theorems:

Lecture 1, Corollary 1.4:

Unions and Intersections of closed sets are closed.

Lecture 1, Definition 1.17 (Compact)

A set S in $\mathbb{R}^n$ is said to be compact if it is **closed** and **bounded**.

Lecture 1, Theorem 1.18 (Weierstrass Theorem):

A continuous function on a **nonempty compact** set $S \subset \mathbb{R}^n$ has a **global maximum point** and a **global minimum point** in S.

Lecture 2, Definition 2.1 (Convex Set)

Used to prove that a set is convex

A set $D \subseteq \mathbb{R}^n$ is said to be convex if for any two points x and y in D, the line segment joining x and y also lies in D. That is,

$$x, y \in D \Rightarrow \lambda x + (1-\lambda)y \in D \quad \forall \lambda \in [0,1]$$

Lecture 2, Proposition 1

Basically intersection of convex sets are also convex

Note that the union on the other hand may not be convex

If C_1, C_2, C_m are convex sets in R^n, then $C = \cap_{i=1}^m C_i$ is also convex.

Lecture 2, Definition 2.6:

Used to prove that a function is convex/concave

Let $D \subseteq \mathbb{R}^n$ be a convex set. Consider a function $f : D \to \mathbb{R}$

(a) The function f is said to be convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \forall x, y \in D, \lambda \in [0,1]$$

(b) The function f is said to be strictly convex if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

For all distinct $x, y \in D,\ \lambda \in (0,1)$

Lecture 2, Proposition 2

Useful properties of convex functions

If $f_1, f_2 : D \to \mathbb{R}$ are convex functions on a convex set $D \subseteq \mathbb{R}^n$, then

(a) $f_1 + f_2$ is a **convex** function on D

(b) αf_1 is a **convex** function on D for $\alpha \geq 0$

(c) αf_1 is a **concave** function on D for $\alpha < 0$

(d) $\max\{f_1, f_2\}$ is a **convex** function on D.

- Note that the $\min(f_1, f_2)$ may not be a convex function

Corollary 2.10:

Let $f_1, f_2, \cdots, f_k : D \to \mathbb{R}$ be convex functions on a convex set $D \subseteq \mathbb{R}^n$. Then

$$f(x) = \sum_{j=1}^{k} \alpha_j f_j(x), \qquad where\ \alpha_j \geq 0, \forall j$$

Is also a convex function on D.

Moreover, if at least one of f_j is strictly convex on D, then f is strictly convex on D.

Lecture 2, Proposition 3:

Useful when we want to show a complicated function is convex.

Let $h: D \to \mathbb{R}$ be a convex function and $g: X \to \mathbb{R}$ be a **non-decreasing convex function** with $h(D) \subset X$.

Then the **composite function** $f = g \circ h: D \to \mathbb{R}$ is a **convex function**.

Lecture 2, Proposition 4:

Useful when we want to show a complicated function is concave.

Let $h: D \to \mathbb{R}$ be a convex function and $g: X \to \mathbb{R}$ be **a non-increasing concave function** with $h(D) \subset X$.

Then the **composite** function $f = g \circ h: D \to \mathbb{R}$ is a **concave function**.

Lecture 2, Proposition 5:

If we have a convex set and convex function, if we can define a set S_α as such then it is a convex set.

Suppose $D \subset \mathbb{R}^n$ is convex. If $f: D \to \mathbb{R}$ is convex, then for any $\alpha \in \mathbb{R}$, the set

$$S_\alpha := \{x \in D | f(x) \leq \alpha\} \text{ is convex}$$

Lecture 2, Proposition 6:

It is basically the area above the graph since we are considering all values of α that is greater than the curve. It tells us if the epigraph is convex or not depending on whether f is a convex function

Suppose $f: D \to \mathbb{R}$ is a function defined on the convex set $D \subset \mathbb{R}^n$. The epigraph of f is the following subset of $\mathbb{R}^{n+1}$:

$$E_f = \{[x; \alpha] : x \in D, \alpha \in \mathbb{R}, f(x) \leq \alpha\}$$

The epigraph E_f is a convex set if and only if f is convex

Theorem 2.17 (Tangent Plane Characterisation of convex functions):

Main idea is just that the tangent plane always lie below the surface for a convex function

Suppose f has continuous first partial derivatives on an open convex set S in $\mathbb{R}^n$. Then

(a) The function f is convex if and only if

$$f(x) + \nabla f(x)^T(y-x) \leq f(y) \ \forall x, y \in S$$

(b) The function f is strictly convex if and only if

$$f(x) + \nabla f(x)^T(y-x) < f(y) \ \forall x \neq \ y \in S$$

Theorem 2.19

Optimality condition for a convex minimization problem over a convex set (Proof is through Tangent Plane)

Let $f : C \to \mathbb{R}$ be a convex and continuously differentiable function on a convex set $C \subset \mathbb{R}^n$. Then $x^* \in C$ is a global minimizer of the minimization problem $\min\{f(x) | x \in C\}$

If and only if

$$\nabla f(x^*)^T(x - x^*) \geq 0, \qquad \forall x \in C$$

Ways to test for definiteness

Note that they need to be square matrices

(1) Definition 3.6: $x^T A x$

Let A be a real $n \times n$ matrix

(a) A is **positive semidefinite** if $x^T A x \geq 0, \forall x \in \mathbb{R}^n$

(b) A is **positive definite** if $x^T A x > 0, \forall x \neq 0$

(c) A is **negative semidefinite** if $x^T A x \leq 0, \forall x \in \mathbb{R}^n$. i.e. $-A$ is positive semidefinite

(d) A is **negative definite** if $x^T A x < 0, \forall x \neq 0$. i.e. $-A$ is positive definite

(e) A is **indefinite** if A is neither positive nor negative semidefinite.

(2) Theorem 3.8 (Eigenvalue Test)

Useful Property: Diagonals of a diagonal matrix are the eigenvalues of the matrix

Let A be a real symmetric $n \times n$ matrix

(a) A is said to be **positive semidefinite** if and only if every eigenvalue of A is **nonnegative** $(\lambda \geq 0)$

(b) A is said to be **positive definite** if and only if every eigenvalue of A is **positive** $(\lambda > 0)$

(c) A is said to be **negative semidefinite** if and only if every eigenvalue of A is **nonpositive** $(\lambda \leq 0)$

(d) A is said to be **negative definite** if and only if every eigenvalue of A is **negative** $(\lambda < 0)$

(e) A is said to be **indefinite** if and only if there is a **positive eigenvalue** of A and a **negative eigenvalue** of A

(3) Theorem 3.11 (Principal Minor Test): Only for positive definite and negative definite

(a) A is **positive definite** if and only if $\Delta_k > 0$ for all $k = 1, 2, \cdots, n$

(b) A is **negative definite** if and only if $(-1)^k \Delta_k > 0$ for all $k = 1, 2, \cdots, n$ (i.e. the principal minors alternate in signs with $\Delta_1 < 0$)

Useful Properties of Eigenvalues:

1) Diagonals of diagonal matrix are eigenvalues of the matrix

2) $\det(A) = \lambda_1 \lambda_2$ (Determinant is product of eigenvalues, useful for 2x2 matrix)

3) $\det(A - \lambda I) = 0$ (Solution to characteristic polynomial are the eigenvalues)

4) Inverse of a 2 x 2 matrix

$$A^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Lecture 3, Definition 3.17 (Coercive function)

A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be coercive if

$$\lim_{||x|| \to \infty} f(x) = +\infty$$

More formally,

$\forall M > 0, \exists r > 0$ such that $||x|| > r \Rightarrow f(x) > M$

To prove coercive: Just make sure that for each component of x, for e.g $x_1, x_2, \cdots, \to +\infty \ \& -\infty, f(x) \to +\infty$, once we prove that then it is coercive.

More formally: Use $||x||_\infty = \max\{|x_1|, \cdots, |x_n|\}$

Use this to show that $f(x) \geq some\ term\ of\ ||x||_\infty$

Once we show this, then we can see that:

$$||x||_\infty \leq ||x|| \leq \sqrt{n}||x||_\infty$$

$$||x|| \to \infty \Leftrightarrow ||x||_\infty \to \infty \Rightarrow f(x) \to \infty$$

Theorem 3.20:

Existence of global min if continuous coercive function

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. If f is a coercive, then f has at least one global minimiser

Theorem 3.22 (Necessary Condition of a local minimiser or local maximiser)

Local Min ⇒ Critical Point, H_f is p.s.d.

Let χ be an open subset of $\mathbb{R}^n$. Suppose $f: \chi \to \mathbb{R}$ has continuous first order partial derivatives in χ.

1) If $x^* \in \chi$ is a local minimiser of f on χ, then x^* is a stationary point, i.e. $\nabla f(x^*) = 0$

In addition, if f has continuous second partial derivatives, then $H_f(x^*)$ is p.s.d

Corollary 3.24

If the Hessian is indefinite it is a saddle point

Let $x^* \in \chi$ be a stationary point of f. If $H_f(x^*)$ is indefinite, then x^* is a saddle point

Theorem 4.7 (Sufficient Condition of local optimizer)

Critical Point, H_f p.d. ⇒ Strict Local minimizer

Let χ be an open subset of $\mathbb{R}^n$. Suppose $f : \chi \to \mathbb{R}$ has continuous second partial derivatives

1) If $x^* \in \chi$ is a stationary point (i.e. $\nabla f(x^*) = 0$) and $H_f(x^*)$ is positive definite, then x^* is a strict local minimiser.

Theorem 4.10

Convex Optimisation Problem then local min is also the global min

Let D be a nonempty open convex subset of $\mathbb{R}^n$, and $f: D \to \mathbb{R}$ is a convex function.

Suppose $x^* \in D$ is a local minimiser to the problem. Then

1) x^* is a global minimiser

2) If f is strictly convex, then x^* is the unique global minimiser.

Corollary 4.11

Stationary point is global min/global max if we have the following conditions

If a f is a convex (respectively concave) function with continuous first partial derivatives on some open convex set D, then any stationary point of f is a global minimiser (respectively maximiser) of f.

Theorem 4.14

Note that Q is the Hessian after we differentiate twice and if it is positive semidefinite then we have a convex function

Let Q be an $n \times n$ symmetric matrix and $c \in \mathbb{R}^n$. The quadratic function $q: \mathbb{R}^n \to \mathbb{R}$ defined by

$$q(x) = \frac{1}{2}x^T Q x + c^T x$$

Is a convex function if and only if Q is p.s.d

Theorem 4.15 (Unconstrained convex quadratic program)

If we have a convex quadratic program, the minimiser is given by the below condition.

Let $c \in \mathbb{R}^n$ and Q be a positive semidefinite matrix. Consider the quadratic function $q: D \to \mathbb{R}$ defined by

$$q(x) = \frac{1}{2}x^T Q x + c^T x$$

Where D is an open convex set of $\mathbb{R}^n$. The point $x^* \in D$ is a global minimiser of q if and only if $Qx^* = -c$

Moreover, if Q^{-1} exists, then $x^* = \ -Q^{-1}c$

(Univariate) Bisection Search (Gradient Method)

(Look at f' not f)

Theorem 4.17 (Intermediate Value Theorem)

Let f' be a continuous function on $[a, b]$, satisfying $f'(a)f'(b) < 0$. Then f' has a root between a and b, that is, there exists a number r satisfying $a < r < b$ and $f'(r) = 0$

Algorithm:

```
1.  Choose interval $[a_1, b_1]$ so that $f'(a_1)$ and $f'(b_1)$ have opposite signs
2.  For $k = 1, 2, \cdots$
    a.  Set $x_k = \frac{1}{2}(a_k + b_k)$
    b.  If $b_k - a_k \leq 2\epsilon$; Stop and use $x_k \in [a_k, b_k]$ as an approximate solution. Else set $[a_{k+1}, b_{k+1}]$ to be $[a_k, x_k]$, $[x_k, b_k]$ choosing the one where the derivative have opposite signs
```

Analysis:

1. $|b_k - a_k| = \frac{|b_1 - a_1|}{2^{k-1}}$
2. At termination, $|b_k - a_k| < \ 2\epsilon$

(Univariate, Multivariate) Newton's Method (Gradient Method)

- Solving for global minimizer of the quadratic approximation of f
- Normally fastest since it is quadratic method
- For quadratic function $q(x) = \alpha x^2 + \beta x + y$, solution is $x^* = \ -\frac{\beta}{2\alpha}$

Newton's Iterate:

Univariate	Multivariate
$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$	$x_{k+1} = x_k - \alpha_k H_f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$

(Multivariate) Solving for optimal step size through Exact Line Search:

$$\alpha_k = \arg\min_{\alpha \geq 0} f\left(x_k - \alpha_k H_f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)\right)$$

Algorithm:

```
1.  Select initial point $x_0$, and TOL $\epsilon > 0$
2.  For $k = 1, 2, \cdots$
    a.  If $|f'(x_k)| < \epsilon$, stop and report $x_k$ as the approximate stationary point
    b.  Else, compute Newton's direction and compute step length by exact line search of Armijo. Set $x_{k+1} = x_k - \alpha_k H_f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right)$
```

Analysis:

- After each iterate, the degree of precision away from optimal solution is doubled. If we can find the distance from optimal, we can approximate the number of iterations needed

(Univariate) Golden Section Search (Non-Gradient Method)

Definition 4.21 (Unimodal Function)

Left side of the global minimiser is strictly decreasing and the right side is strictly increasing. Means that there is exactly one global minimizer. Used for Golden Section Search

A function f is unimodal on $[a, b]$ if it has exactly one global minimiser in the interval $[a, b]$, and it is strictly decreasing on $[a, x^*]$ and strictly increasing on $[x^*, b]$.

Algorithm:

```
1.  Set $[a_0, b_0] = [a, b]$. Choose $\epsilon > 0, \alpha = \frac{\sqrt{5}-1}{2}$. Let $\lambda_0 = b - \alpha(b - a), \mu_0 = a + \alpha(b - a)$
2.  Evaluate $f(\lambda_0), f(\mu_0)$
3.  For $k = 0, 1, 2, \cdots$
```

If $f(\lambda_k) > f(\mu_k)$: $\alpha_{k+1} = \lambda_k, \quad b_{k+1} = b_k$ $\lambda_{k+1} = \mu_k,$ $\mu_{k+1} = \lambda_k + \alpha(b - \lambda_k)$ Compute $f(\mu_{k+1})$	If $f(\mu_k) \geq f(\lambda_k)$: $\alpha_{k+1} = a_k, \quad b_{k+1} = \mu_k$ $\mu_{k+1} = \lambda_k,$ $\lambda_{k+1} = \mu_k - \alpha(\mu_k - a)$ Compute $f(\lambda_{k+1})$

Analysis:

The range shrinks to $\alpha^n(b_0 - a_0)$ at the nth iteration.

Armijo Line Search:

- Fast but may not find the smallest α_k
- $p^{(k)}$ direction of descent, σ – Indicator of whether the functional value is small enough, β – Shrinkage of α value

Algorithm:

```
1.  Let $\sigma \in (0, 0.05)$ and $\beta \in (0, 1)$. Choose an initial step length $\bar{\alpha}$
2.  For $r = 1, 2, \cdots$ do
    a.  Set $\alpha = \beta^r \bar{\alpha}$
    b.  If $f\left(x^{(k)} + \alpha p^{(k)}\right) \leq f\left(x^{(k)}\right) + \alpha\sigma\nabla f\left(x^{(k)}\right)^T p^{(k)}$ then break
```

Required step length $\alpha = \beta^r \bar{\alpha}$

(Multivariate) Steepest Descent Algorithm (Gradient Method)

Algorithm:

```
1.  Select an initial point $x^{(0)}, \epsilon > 0$
2.  For $k = 0, 1, 2, \cdots$
    a.  Evaluate $d^{(k)} = -\nabla f\left(x^{(k)}\right)$
    b.  If $\left|\left|d^{(k)}\right|\right| < \epsilon$, stop the algorithm; $x^{(k)}$ is an approximate solution.
    c.  Else, find the value of $t_K$ that minimizes the one-dimensional function
            $g(t) := f\left(x^{(k)} + td^{(k)}\right)\ over\ t \geq 0$
        Set $x^{(k+1)} = x^{(k)} + t_k d^{(k)}$
```

Monotonic Decreasing Property:

If $x^{(k)}$ is a steepest descent sequence for a function $f(x)$, and if $\nabla f\left(x^{(k)}\right) \neq 0$ for some k, then $f(x^{k+1}) < f(x^k)$

Convergence of Steepest Descent:

If $f(x)$ is a coercive function, then the limit of any convergent subsequence of $\left\{x^{(k)}\right\}$ is a critical point of $f(x)$

Analysis:

For convex quadratic optimization problem **Convergence Rate**: When $\kappa(Q)$ is large

$$\kappa(Q) = \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)}, \qquad \rho(Q) = 1 - \frac{4}{\kappa(Q)}$$

(Multivariate) Coordinate Descent Algorithm

- Good when we have large problems

Algorithm:

```
1.  Specify some initial guess of $x^{(0)}$
2.  For $k = 0, \cdots$
    a.  If $x^{(k)}$ is optimal then stop
    b.  Else for $i = 1,2, \cdots, n$
            $x^{(k+1)} = \arg\min_{x_i \in \mathbb{R}} f\left(x_i, \omega_{-i}^{(k)}\right)$
```

Jacobi Rule:

- Doesn't use the most updated values, just use values from the previous iteration
- Easily parallelizable

Gauss-Seidel Rule:

- Makes use of the most updated values
- Hard to be parallelized

Update for Linear Regression:

$$x_p^{(k+1)} = \frac{A_p^T r^{(p,k)}}{A_p^T A_p} + x_p^{(k)}$$

$$r^{(1,k)} = b - Ax^{(k)}, \qquad r^{(p,k)} = r^{(p-1,k)} + \left(x_{p-1}^{(k)} - x_{p-1}^{(k+1)}\right)A_{p-1}$$

(Multivariate) Stochastic Gradient Descent Algorithm

$$f(x) = E(g(x, a), z)$$

$$L\left(g\left(x^{(k)}, a^{(k)}\right), z^{(k)}\right) = g\left(x^{(k)}, a^{(k)}\right), z^{(k)}$$

Algorithm:

Suppose that we have data $z_1 = (a_1, b_1), \cdots, z_n = (a_n, b_n)$

```
1.  Pick an initial point $x^{(0)}$
2.  Find a step size sequence $t_k$
3.  Repeat the following:
    a.  Draw a random sample $z^{(k)}$ from $\{z_1, \cdots, z_n\}$ and update
            $x^{(k+1)} = \ x^{(k)} - t_k \nabla L\left(x^{(k)}, z^{(k)}\right)$
4.  Output the final $x^{(k+1)}$
```

Constrained Problem Theorems:

Definition 7.3 (Linear Independence Constraint Qualification (LICQ)

Find Regular Point

- Get the set of equality constraints and active inequality constraints (i.e. $h(x) = 0$)
  $\{\nabla g_i(x^*) : i = 1, \cdots, m\} \cup \{\nabla h_j(x^*) : j \in J(x^*)\}$
- Check that above set of vectors are linearly independent
  - If they are linearly independent then x^* is a regular point. Else x^* is not a regular point
  - Only 1 vector in the set and it is the 0 vector → Linearly dependent → Not regular
  - More columns than rows → Linearly dependent → Not regular
  - Interior Points → Regular Points by definition

Definition 8.2 (KKT First Order Necessary Condition):

Suppose x^* is a regular point, x^* satisfies the KKT first order (necessary) conditions if

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g(x^*) + \sum_{j=1}^{p} \mu_j^* \nabla h_j(x^*) = 0$$

$$\lambda_i^* \in \mathbb{R}, \qquad \mu_j^* \ge 0, \forall j = 1, 2, \cdots, p, \qquad \mu_j^* = 0 \ \forall j \notin J(x^*)$$

Where $J(x^*)$ is the index set of active inequality constraints at x^*

Complementary Slackness:

$$\mu_j^* h_j(x^*) = 0, \qquad \forall j = 1, 2, \cdots, p$$

- Lagrange multiplier for inactive inequality constraints is 0. i.e $\mu_j = 0$ when $h_j(x^*) < 0$
- Lagrange multiplier can be non-zero for active inequality constraints. i.e. $h_j(x^*) = 0 \ \forall j \in J(x^*)$ and $\mu_j^* \ge 0$

Definition 8.6 (KKT Second Order Necessary Conditions):

x^* satisfies the KKT second order (necessary conditions) if

1. x^* is a KKT point (it satisfies the KKT first order necessary conditions)
2. $y^T H_L(x^*) y \ge 0$, for all $y \in C(x^*, \lambda^*, \mu^*)$

$$H_L(x^*) = H_f(x^*) + \sum_{i=1}^{m} \lambda_i^* H_{g_i}(x^*) + \sum_{j=1}^{p} \mu_j^* H_{h_j}(x^*)$$

$C(x^*, \lambda^*, \mu^*)$:

$$= \left\{ y \in \mathbb{R}^n : \begin{array}{ll} \nabla g_i(x^*)^T y = 0, & i = 1, 2, \cdots, m \\ \nabla h_j(x^*)^T y = 0\,, & j \in J(x^*) \text{ and } \mu_j > 0 \\ \nabla h_j(x^*)^T y \le 0\,, & j \in J(x^*) \text{ and } \mu_j = 0 \end{array} \right.$$

Theorem 8.6 (KKT Necessary Conditions)

- f, g_i, h_j has continuous first partial derivatives on the feasible set S
- $x^* \in S$ is regular

First order necessary condition:

- If x^* is a local minimizer, then x^* is a KKT point

Second order necessary condition:

- If f, g_i, h_j has continuous second partial derivatives on the feasible set S, then x^* also satisfies the KKT second order necessary conditions

Corollary 8.8:

- With the conditions in Theorem 8.6:
- If x^* is a global minimizer, then x^* is a KKT point
- If x^* is not a KKT point, x^* is not a global minimizer

Lecture 8, Proposition 1 (Easier way to check Definiteness):

- Strict Complementarity holds at x^* (i.e. $\mu_j > 0$ if $j \in J(x^*)$)

We can consider the matrix

$$\mathcal{D}(x^*) = \left(\nabla g_1(x^*), \cdots, \nabla g_m(x^*), [\nabla h_j(x^*) : j \in J(x^*)]\right)$$
$$Z(x^*) = \{x \in \mathbb{R} : \mathcal{D}(x^*)^T x = 0\}$$

$$y^T H_L(x^*) y \ge 0 \ \ \forall y \in C(x^*, \lambda^*, \mu^*)$$
$$\Leftrightarrow Z(x^*)^T H_L(x^*) Z(x^*) \ \ is\ p.s.d.$$

Note that we can use this to check the definiteness of $H_L(x^*)$ instead of finding y from the critical cone

- If $H_L(x^*)$ is positive definite, then $Z(x^*)^T H_L(x^*) Z(x^*)$ is also **positive definite**

Theorem 8.12 (KKT Sufficient Condition):

KKT point + H_L p.d. → strict local minimizer

- f, g_i, h_j be functions with continuous first and second derivatives
- Suppose $x^* \in S$ is a KKT point (First Order KKT necessary conditions met)
  $$y^T H_L(x^*) y > 0\,, \qquad \forall y \in C(x^*, \lambda^*, \mu^*)$$
- Then x^* is a **strict local minimizer** of f on S if $H_L(x^*)$ is p.d.

Theorem 9.2 (KKT Point is Optimal Solution Under Convexity):

Convex Program, KKT point ⇒ global min

- f, h_j are differentiable convex functions
- $g_i := a_i^T x - b$ which means it is linear

If $x^* \in S$ is a KKT point, then x^* is a global min of f on S.

Slater's Condition:

Find a point where equality constraint is satisfied, and inequality constraint is not active

There exists $\hat{x} \in \mathbb{R}^n$ such that $g_i(\hat{x}) = 0, \forall\, i = 1, \cdots, m$ and $h_j(\hat{x}) < 0 \ \ \forall j = 1, \cdots, p$.

Theorem 9.5 (Optimal Solution is KKT point):

Convex Program, global min, Slater's Condition hold ⇒ KKT Point

- f, h_j are differentiable convex functions
- $g_i := a_i^T x - b$ which means it is linear
- At least 1 inequality constraint
- Slater's condition holds (If no inequality constraint, this immediately holds)

If $x^* \in S$ is a global minimizer on S, then x^* is a KKT point

Theorem 9.7 (Linear Equality Constrained NLP (ECP))

$$\min f(x)$$
$$s.t. Ax = b, \qquad x \in \mathbb{R}^n$$

- A is a $m \times n$ matrix whose rows $\{a_i^T\}_{i=1}^m$ are linearly independent. (Regularity Condition)
- f is differentiable convex function. Note that (⇐) we don't need convexity of f.
- $x \in S^*$ is a KKT point ⇔ x^* is a global minimizer of f

Lecture 9, Proposition 1 (Perturbation of $F(c)$ with respect to changes in constraints)

$$\frac{\partial F(c)}{\partial c_k} = \frac{\partial f(x^*(c))}{\partial c_k} = \lambda_k^*(c), \forall k = 1, \cdots, m$$

- A small change in the kth constraint from $g_k(x) = 0$ to $g_k(x) + c_k = 0$. The new optimal objective value is ≈ $f(x^*) + \lambda_k^* c_k$

Lagrangian Function:

$$L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x)$$

Lagrangian Dual Function:

$$\theta(\lambda, \mu) = \inf_{x \in X} L(x, \lambda, \mu) = \inf_{\mathrm{x}} \{f(x) + \lambda^T g(x) + \mu^T h(x) | x \in X\}$$

Lagrangian Dual Problem:

$$\max_{\lambda \in \mathbb{R}^m, \mu > 0} \theta(\lambda, \mu) = \max_{\lambda \in \mathbb{R}^m, \mu > 0} \inf_{\mathrm{x}} \{f(x) + \lambda^T g(x) + \mu^T h(x) | x \in X\}$$

Lecture 10, Proposition 3 (Concavity of Lagrangian Dual Function):

If $\theta(\lambda, \mu) = \inf_{x \in X} L(x, \lambda, \mu) = \inf_{\mathrm{x}} \{f(x) + \lambda^T g(x) + \mu^T h(x) | x \in X\}$ is finite for all (λ, μ) with $\mu \ge 0$ then $\theta(\lambda, \mu)$ is a **concave** function.

Theorem 10.7 (Weak Duality Theorem):

Let x be a feasible solution to (P) and (λ, μ) be a feasible solution to (D).

$$f(x) \ge \theta(\lambda, \mu)$$

Corollary 10.9 (Using Theorem 10.7)

- Optimal primal (minimization) objective value ≥ Optimal dual (maximization) objective value
  $$\min\{f(x) : x \in S\} \ge \max\,\{\theta(\lambda, \mu) : \lambda \in \mathbb{R}^m, \mu \ge 0\}$$
- If x^* is a feasible solution to (P) and (λ^*, μ^*) is a feasible solution to (D) such that
  $$f(x^*) = \theta(\lambda^*, \mu^*)$$

Then **x^* is an optimal solution to (P) and (λ^*, μ^*) is an optimal solution to (D)**. Makes the first part of the Corollary all equality

Theorem 10.12 (Strong Duality Theorem):

- X is a convex set, f, h_j are convex functions, g_i are affine functions
- Slater's Condition hold
- Then duality gap is 0
  $$\inf\{f(x) : x \in S\} = \sup\,\{\theta(\lambda, \mu) : \lambda \in \mathbb{R}^m, \mu \ge 0\}$$
- Also if inf in (P) is finite, then sup is attained at some (λ_*, μ_*). If inf is attained at x^*, then $\mu_*^T h(x^*) = 0$

Subgradient Descent/Ascent Method:

- Can be used when we are cannot differentiate f

Definition 11.2 (Subgradient):

- S nonempty convex set
- f is a convex function

A vector $\xi \in \mathbb{R}^n$ is a subgradient of f at $\bar{x} \in S$ if

$$f(x) \ge f(\bar{x}) + \xi^T (x - \bar{x}), \qquad \forall x \in S$$

Subdifferential of f at $\bar{x}$ is the set of all subgradients of f at $\bar{x}$

$$\partial f(\bar{x}) = \{\xi : \xi \text{ is a subgradient of } f \text{ at } \bar{x}\}$$

Lecture 11 Propositions for Subgradient:

Proposition 1:

If f is differentiable at x, then

$$\partial f(x) = \{\nabla f(x)\}$$

Proposition 2:

If f is continuous and convex

$\min_{x \in \mathbb{R}^n} f(x)$ is attained at x^* ⇔ $0 \in \partial f(x^*)$

Proposition 3:

The subdifferential of $f + g$ is given by:

$$\partial(f + g)(x) \supseteq \{u + v\, | u \in \partial f(x), v \in \partial g(x)\}$$

- Basically the addition of the all combinations of the subdifferentials

Proposition 4:

If $S = \{v_1, \cdots, v_n\}$ then

$$conv(S) = \left\{ v = \sum_{i=1}^{n} \lambda_i v_i\,, \qquad \lambda_i \ge 0\,, \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

Proposition 5:

Suppose $f(x) = \max\{f_1(x), \cdots, f_m(x)\}$ where f_i are all convex and continuously differentiable functions. If $f(x^*) = f_1(x^*) = \cdots = f_j(x^*)$

$$\partial f(x^*) = conv\left(\{\nabla f_1(x^*), \cdots, \nabla f_j(x^*)\}\right)$$

- Useful for $|x|$ type of functions

Algorithm:

```
1.  Specify some initial guess of $x^{(0)}$
2.  For $k = 0, 1, \cdots$
    a.  If $0 \in \partial f\left(x^{(k)}\right)$, thenstop
    b.  Else, pick $v^{(k)} \in -\partial f\left(x^{(k)}\right)$. Set
            $x^{(k+1)} = x^{(k)} + t_k v^{(k)}$
```

Take last $x^{(k+1)}$ as minimizer

Projected Gradient Descent:

Theorem 11.8 (Projection Theorem):

Let C be a closed and convex set in $\mathbb{R}^n$

(a) For every $z \in \mathbb{R}^n$, there exists a unique minimizer for the projection of z onto C

$$\Pi_C(z) = \arg\min_{z} \left\{ \frac{1}{2} ||x - z||^2 \,\middle|\, x \in C \right\}$$

(b) $x^* := \Pi_C(z)$ is the projection of z onto C if and only if

$$\langle z - x^*, x - x^* \rangle \le 0, \qquad \forall x \in C$$

(c) For any $z, w \in \mathbb{R}^n$

$$||\Pi_C(z) - \Pi_C(w)|| \le ||z - w||$$

(d) If C is a linear subspace of $\mathbb{R}^n$, then $(z - x^*) \perp C$. Therefore, z can be decomposed into two perpendicular components:

$$z = \Pi_C(z) + \left(z - \Pi_C(z)\right)$$
$$\langle z - \Pi_C(z), \Pi_C(z) \rangle = 0$$

(e) If C is a closed convex cone, then it is also true that

$$\langle z - \Pi_C(z), \Pi_C(z) \rangle = 0$$

Cone: A set $\Omega \subset \mathbb{R}^n$ is said to be a cone if $\lambda x \in \Omega$, whenever $x \in \Omega$, and $\lambda \ge 0$

To find the projection Π_C:

Solve the minimization problem

$$\min_x \left\{ \frac{1}{2} ||x - y||^2 \right\} \ \ s.t. x \in C$$

Solve it via the KKT system since it is a constrained problem.

$$\Pi_C(y) = \begin{cases} y, & if\ y \in C \\ KKT\ Solution, & if\ y \notin C \end{cases}$$

Algorithm:

```
1.  Select an initial point $x^{(0)}, \epsilon > 0$
2.  For $k = 0, 1, 2, \cdots$
    a.  Evaluate $d^{(k)} = -\nabla f\left(x^{(k)}\right)$
    b.  If $\left|\left|x^{(k+1)} - x^{(k)}\right|\right| < \epsilon$, stop the algorithm; $x^{(k)}$ is
        an approximate solution.
    c.  Else, find the value of $t_K$ that minimizes the one-
        dimensional function
            $g(t) := f\left(x^{(k)} + t d^{(k)}\right)\ over\ t \ge 0$
        Set $x^{(k+1)} = \Pi_S\left(x^{(k)} + t_k d^{(k)}\right)$
```

Common Projections:

$$S = \{||x|| \le 1\}$$

$$\Pi_C(y) = \begin{cases} y, & if\ ||y|| \le 1 \\ \dfrac{y}{||y||}, & otherwise \end{cases}$$

$$S = \{a^T x + b \le 0\}$$

$$\Pi_C(y) = \begin{cases} y, & if\ a^T y + b \le 0 \\ y - \dfrac{a^T y + b}{||a||^2} a, & otherwise \end{cases}$$

Quadratic Penalty Method:

- For equality constrained NLP
  $$\min f(x)\,, \qquad s.t. c_i(x) = 0, i \in \mathcal{E}$$
- Issue is that when $\mu \to 0$, H_Q^{-1} can be very singular, which can give numerical problems.
- We want to get $\mu \to 0$ so that the constraint of $c_i(x) = 0$.
- Normally, we can solve it like a unconstrained problem, so we just find the stationary points.

Quadratic Penalty Function:

$$Q(x; \mu) = f(x) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i^2(x)$$

Gauss-Newton Approximation:

$$H_Q(x, \mu) \approx H_f(x) + \frac{1}{\mu} \sum_{i \in \mathcal{E}} \nabla c_I(x)^T \nabla c_i(x)$$

Algorithm:

```
1.  Choose a starting point $x^{(0)}$ and stopping tolerance $\epsilon$.
    Set $\mu_0 = 1$.
2.  For $k = 0, 1, \cdots$
    a.  Find an approximate minimizer $x^{(k+1)}$ of $Q(x; \mu_k)$
        (e.g. using Newton's method and taking $x^{(k)}$ as
        initial guess)
    b.  Stop if $\left|\left|c\left(x^{(k+1)}\right)\right|\right| < \epsilon$
    c.  Else choose new $\mu_{k+1} = \rho \mu_k,\ \rho < 1$
```

Final Convergence Test can also be:

$$\left|\left| \nabla f\left(x^{(k+1)}\right) + \sum_{i \in \mathcal{E}} \lambda_i^{(k)} \nabla c_i\left(x^{(k+1)}\right) \right|\right| < \epsilon$$

Where $\lambda_i^{(k)} = c_i(x^{(k+1)}) / \mu_k$

Augmented Lagrangian Method:

- For **equality** constrained NLP
- Exact penalty method, does not need $\mu \downarrow 0$
- Solve it like a constrained problem with KKT

Augmented Lagrangian:

$$\min_x L_A(x, \lambda, \mu) := f(x) + \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \frac{1}{2\mu} \sum_{i \in \mathcal{E}} c_i(x)^2$$

Optimality Condition for Augmented Lagrangian:

$$\nabla f\left(x_A^{(k+1)}\right) + \sum_{i \in \mathcal{E}} \left[\left[\lambda_i^{(k)} + \frac{c_i\left(x_A^{(k+1)}\right)}{\mu_k} \right] \nabla c_i\left(x_A^{(k+1)}\right) \right] = 0$$

This is like the normal Lagrangian where

$$\lambda^{(k+1)} = \lambda_i^{(k)} + \frac{c_i\left(x_A^{(k+1)}\right)}{\mu_k}$$

Algorithm:

```
1.  Choose $\mu_0 > 0, \tau_0 > 0$. Choose starting points $x^{(0)}, \lambda^{(0)}$
2.  For $k = 0, 1, 2, \cdots$
    a.  Find an approximate minimizer $x^{(k+1)}$ of
        $L_A(x, \lambda, \mu)$ (e.g. using Newton's method and taking
        $x^{(k)}$ as initial guess)
    b.  If final convergence test satisfied, stop
    c.  Else, set
            $\lambda^{(k+1)} = \lambda_i^{(k)} + \frac{c_i\left(x^{(k+1)}\right)}{\mu_k}$
        Choose new $\mu_{k+1}, \tau_{k+1}$
```

Barrier Function Methods:

- For **inequality** constraints. Assuming that f is continuously differentiable
  $$\min f(x)\,, \qquad s.t.\ c_i(x) \le 0, i \in \mathcal{E}$$

Barrier Function:

$$B(x) = \sum_{i \in I} \phi\left(-c_i(x)\right), \qquad where\ \phi : \mathbb{R}_+ \to \mathbb{R}_+$$

- $\phi'(y) < 0$ (ϕ is strictly decreasing)
- $\lim_{y \to 0^+} \phi(y) = \infty$ (Close to boundary is penalized)

Example: $\phi = -\log(.)$

Barrier Problem:

$$\min P(x, \mu_k) := f(x) + \mu_k B(x)$$
$$s.t.\ \ c_i(x) < 0$$
$$F^< := \{x \in \mathbb{R}^n : c_i(x) < 0, i \in I\}$$

Note that $c_i(x)$ now has strict inequality and we consider $\mu > 0$

- We can solve it like a normal unconstrained problem, finding the stationary point.

Algorithm:

```
1.  Choose a $\mu_0 > 0, \tau_0 > 0$, starting point $x^{(0)}$
2.  For $k = 0, 1, \cdots$
    a.  Find an approximate minimizer $x^{(k+1)}$ of $P\left(x^{(k+1)},
        \mu_k\right)$ (e.g. using Newton's method and taking $x^{(k)}$ as
        initial guess)
    b.  If final convergence test satisfied, stop
    c.  Else choose new $\mu_{k+1} \in (0, \mu_k),\ \tau_{k+1}$
```

Dot Product Properties:

$$\langle a, b \rangle \Leftrightarrow a \cdot b$$

$$\langle a, b \rangle = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + \cdots + a_n b_n$$

$$||a|| = \sqrt{a_1^2 + \cdots + a_n^2}$$

$$\langle a, b \rangle = ||a||\,||b|| cos\theta$$

$$||a|| = \sqrt{\langle a, a \rangle} =$$

1. **Commutative**: $\langle a, b \rangle = \langle b, a \rangle$
2. **Distributivity**: $\langle a, b + c \rangle = \langle a, b \rangle + \langle a, c \rangle$
3. **Bilinear**: $\langle a, (rb + c) \rangle = r\langle a, b \rangle + \langle a, c \rangle$
4. **Scalar Multiplication**: $\langle c_1 a, c_2 b \rangle = c_1 c_2 \langle a, b \rangle$
5. **Not Associative**
6. **Orthogonal**: Two non-zero vectors are orthogonal if and only if $\langle a, b \rangle = 0$
7. **No Cancellation**: For $\langle a, b \rangle = \langle a, c \rangle$ and $a \ne 0$, we cannot just make it $\langle b \rangle = \langle c \rangle$
8. **Product Rule**: If a and b are differentiable functions, then the derivative: $\langle a, b \rangle' = \langle a', b \rangle + \langle a, b' \rangle$