Last Class : July 21
Final Exam : July 24 1:30~4:30 pm

MAT3007 Optimization
Lecture 17 KKT Conditions
Algorithms

Yuang Chen

School of Data Scienc
The Chinese University of Hong Kong, Shenzhen

July 14, 2025

Outline

Outline

Lagrangian Function

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & g_i(x) \leq b_i, \quad \forall i \in I \\
& h_j(x) = d_j, \quad \forall j \in J
\end{aligned}
\qquad \mathcal{X}
$$

The Lagrangian function is

$$L(x, \lambda, \mu) = f(x) + \sum_{i \in I} \underbrace{\underbrace{\lambda_i}_{\geq 0} \underbrace{[g_i(x) - b_i]}_{\leq 0}}_{\leq 0} + \sum_{j \in J} \underbrace{\mu_j \underbrace{[h_j(x) - d_j]}_{=0}}_{=0} \quad \leq f(x) \quad \forall x \in \mathcal{X}$$

where $\lambda_i \geq 0$ for all $i \in I$ and μ_j for $j \in J$ are Lagrangian multipliers.

Lagrangian Relaxation

Original Problem:

$$(P): \quad v_P := \min\{f(x) : g_i(x) \le b_i, \forall i \in I, \; h_j(x) = d_j, \forall j \in J\}$$

For $\lambda \ge 0$, the Lagrangian relaxation is

$$\begin{aligned}\mathcal{L}(\lambda, \mu) &:= \min_x L(x, \lambda, \mu) \\ &= \min_x \left\{ f(x) + \sum_{i \in I} \lambda_i [g_i(x) - b_i] + \sum_{j \in J} \mu_j [h_j(x) - d_j] \right\}.\end{aligned}$$

Then:

$$\mathcal{L}(\lambda, \mu) \le v_P \quad \forall \lambda \ge 0.$$

Lagrangian Dual Problem

Original Problem:

$$(P): \quad v_P = \min_x \{f(x) : g_i(x) \le b_i, \ \forall i \in I, \ h_j(x) = d_j, \ \forall j \in J\}$$

(Lagrangian) Dual Problem:

$$\begin{aligned}(D): \quad v_D &:= \max_{\lambda \ge 0, \ \mu} \mathcal{L}(\lambda, \mu) \\ &= \max_{\lambda \ge 0, \ \mu} \min_x L(x, \lambda, \mu) \\ &= \max_{\lambda \ge 0, \ \mu} \min_x \left\{ f(x) + \sum_{i \in I} \lambda_i [g_i(x) - b_i] + \sum_{j \in J} \mu_j [h_j(x) - d_j] \right\}\end{aligned}$$

Weak duality: $v_D \le v_P$

Weak Duality

Theorem (Weak Duality)

If the primal and dual optimal value is v_P and v_D, then

$$v_D \leq v_P$$

Note that this always holds (even if the primal problem is nonconvex).

Duality Gap $= V_p - V_D$

Strong Duality : $V_p = V_D$, duality gap $= 0$.

Strong Duality Theorem

Theorem (Strong Duality Theorem)

① ②

If the primal problem is a convex optimization problem and also Slater's condition holds, then strong duality holds (duality gap is 0), i.e., $v_D = v_P$.

Definition (Slater's Condition)

If the original problem (P) is convex (i.e., f and g_i are convex for all $i \in I$, and h_j are affine for all $j \in J$), and there exists a strictly feasible $x \in \mathbb{R}^n$ such that:

$$g_i(x) < b_i \ \ \forall i \in I, \quad h_j(x) = d_j \ \ \forall j \in J,$$

then Slater's Condition holds and strong duality holds.

- Actually only need strict inequalities for non-affine h_i
- In other words, Slater's condition requires that there is a strict interior solution for nonlinear constraints.

Outline

General Optimality Conditions

Now, We consider the following nonlinear program:

$L(x,\lambda,\mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$

$$
\begin{aligned}
\min_{\cancel{x \in \mathbb{R}^n}} \quad & f(x) \\
\text{subject to} \quad & g_i(x) \le 0, \quad \forall i = 1, \ldots, m, \quad (\lambda_i) \\
& h_j(x) = 0, \quad \forall j = 1, \ldots, p. \quad (\mu_j)
\end{aligned}
$$

X

We will also work with the following index sets:

Definition: Active and Inactive Set

At a point $x \in X$, the set $\mathcal{A}(x) := \{i : g_i(x) = 0\}$ denotes the set of **active constraints**. The set of **inactive constraints** is then given by $\mathcal{I}(x) := \{i : g_i(x) < 0\}$.

Karush-Kuhn-Tucker Conditions

Karush-Kuhn-Tucker (KKT) Conditions:

$$\partial_x L(x, \lambda, \mu) = 0$$

subgradient

- Main/Stationarity Condition

if f, g, h are differentiable

$$\nabla_x L(x, \lambda, \mu) \stackrel{!}{=} \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{p} \mu_j \nabla h_j(x) = 0$$

- Complementary Slackness

$$\lambda_i \cdot g_i(x) = 0 \quad \forall i = 1, \ldots, m$$

$$\mu_j h_j(x) = 0 \quad \forall j = 1, \cdots, p$$

- Primal Feasibility

$$g_i(x) \leq 0, \quad h_j(x) = 0 \quad \forall i = 1, \ldots, m, \ \forall j = 1, \ldots, p$$

- Dual Feasibility

$$\lambda_i \geq 0 \quad \forall i = 1, \ldots, m$$

Theory with KKT Conditions

Theorem

1. *For any optimization problem with a constraint qualification (CQ) holds (e.g., strong duality), if* $x^\star$ *and* $\lambda^\star, \mu^\star$ *are primal and dual (local) optimal solutions, then* $x^\star$ *and* $\lambda^\star, \mu^\star$ *satisfy the KKT conditions.*
2. *For convex optimization problem, if* $x^\star$ *and* $\lambda^\star, \mu^\star$ *satisfy the KKT conditions, then* $x^\star$ *and* $\lambda^\star, \mu^\star$ *are primal and dual optimal solutions and strong duality holds.*

Convex OPT + Strong duality holds

KKT $\Longleftrightarrow$ optimal solutions

Proof:

1. Statement:

$$\boxed{\begin{array}{c}\text{Strong duality} + x^*, \lambda^*, \mu^* \text{ optimal} \\ \Downarrow \\ x^*, \lambda^*, \mu^* \text{ satisfy KKT conditions}\end{array}}$$

Since strong duality holds, we have

$$V_p = f(x^*) = \mathcal{L}(\lambda^*, \mu^*) = V_D$$

$$= \min_x L(x, \lambda^*, \mu^*)$$

by def of min of a function

$$\overset{①}{\leq} L(x^*, \lambda^*, \mu^*)$$

$$= f(x^*) + \underbrace{\sum_i \underbrace{\lambda_i^*}_{\geq 0} \underbrace{g_i(x^*)}_{\leq 0}}_{\leq 0} + \underbrace{\sum_j \mu_j^* \underbrace{h_j(x^*)}_{=0}}_{=0}$$

because x^*, λ^*, μ^* are primal & dual feasible

$$\overset{②}{\leq} f(x^*)$$

We get $f(x^*) \leq f(x^*)$, so all above '$\leq$' should be "$=$"

For ① '$\leq$', $\min_x L(x, \lambda^*, \mu^*) = L(x^*, \lambda^*, \mu^*)$

$\Rightarrow x^*$ is the minimizer for $L(x, \lambda^*, \mu^*)$

By FONC, $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$

This gives the main condition in KKT.

For ② '$\le$', $\sum_i \lambda_i^* g_i(x^*) = 0$

$$\Rightarrow \quad \lambda_i^* g_i(x^*) = 0, \quad \forall i$$

This gives the complementary slackness in KKT.

2. For convex optimization
x^*, λ^*, μ^* satisfy KKT
$\Downarrow$
x^*, λ^*, μ^* are optimal solns
+ Strong duality holds

From main condition:

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0$$

$$L(x, \lambda^*, \mu^*) = f(x) + \sum_i \lambda_i^* g_i(x) + \sum_j \mu_j^* h_j(x)$$

$L(x, \lambda^*, \mu^*)$ is a convex function in x

because we consider a convex opt Problem.

$\Rightarrow x^*$ is the minimizer of $L(x, \lambda^*, \mu^*)$

For a convex unconstrained optimization Problem
$\min_x L(x)$, assume $L(x)$ is differentiable

Prove if $\nabla L(x^*) = 0$, then x^* is a global min

$$\mathcal{L}(\lambda^*, \mu^*) = \min_{x} L(x, \lambda^*, \mu^*)$$

$$= L(x^*, \lambda^*, \mu^*) \quad \text{from main condition}$$

$$= f(x^*) + \sum_i \cancelto{0}{\lambda_i^* g_i(x^*)} + \sum_i \cancelto{=0}{\mu_j^* h_j(x^*)}$$

Complementary slackness

Primal feasible

$$= f(x^*)$$

Duality gap is 0, thus x^*, λ^*, μ^* are optimal solutions for primal and dual.

→ Strong duality holds.

Constraint Qualifications

We require the collection of gradients

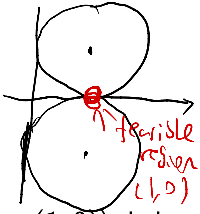
$$\{\nabla g_i(x) : i \in \mathcal{A}(x)\} \cup \{\nabla h_j(x) : j = 1, \ldots, p\} \tag{CQ}$$

to be linearly independent or to have full rank.

- This condition is a **constraint qualification (CQ)** and is called **Linear Independence Constraint Qualification (LICQ)**.
- A feasible point x satisfying the LICQ is called **regular**.
- There are more CQs: ACQ, GCQ, MFCQ, PLICQ, Slater's condition (Optimization II (?)).

Failure of The KKT Conditions in the Absence of CQ

$$
\begin{aligned}
\min_{x_1, x_2} \quad & x_1 \\
\text{subject to} \quad & (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1, \\
& (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1.
\end{aligned}
$$



- Since there is only one feasible solution (i.e., $(x_1, x_2) = (1, 0)$), it is automatically optimal.
- However, this point does not satisfy the KKT conditions (exercise). how?
- This example says if a local min does not hold certain CQ, the KKT conditions may not be necessary at this point.
- This means that points satisfying KKT conditions may not contain all the local minimizers, since some of them may not be regular.

Remarks on KKT Conditions

Remarks:

- KKT conditions is first-order necessary conditions (FONC) for general constrained optimization problems.
- KKT conditions unify all formerly studied FONC.
- A (feasible) point (often the primal and dual variables (x, λ, μ) together) satisfying the KKT conditions is called a KKT point, regardless whether it satisfies CQ or not. We may also call x a KKT point for simplicity.
- KKT points are candidates for local optimal solutions – just like stationary points.

Example 1

Consider the problem:

$$\begin{aligned} \text{minimize} \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 \le 5 \quad (u_1) \\ & 3x_1 + x_2 \ge 3 \quad (u_2) \\ & -3x_1 - x_2 \le -3 \end{aligned}$$

Find the KKT conditions.

$$L(x, u_1, u_2) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 + u_1(x_1^2 + x_2^2 - 5) + u_2(3 - 3x_1 - x_2)$$

$$u_1, u_2 \ge 0$$

KKT Conditions : ⓐ + ⓑ + ⓒ + ⓓ

Main :

$$\nabla_x L(x, u_1, u_2) = \begin{bmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} 4x_1 + 2x_2 - 10 + 2u_1x_1 - 3u_2 \\ 2x_1 + 2x_2 - 10 + 2u_1x_2 - u_2 \end{bmatrix}$$

ⓐ
$$\begin{cases} 4x_1 + 2x_2 - 10 + 2u_1x_1 - 3u_2 = 0 \\ 2x_1 + 2x_2 - 10 + 2u_1x_2 - u_2 = 0 \end{cases}$$

Complementary Slackness

ⓑ
$$\begin{cases} u_1 (x_1^2 + x_2^2 - 5) = 0 \\ u_2 (3 - 3x_1 - x_2) = 0 \end{cases}$$

Primal feasible

ⓒ
$$\begin{cases} x_1^2 + x_2^2 \leq 5 \\ 3x_1 + x_2 \geq 3 \end{cases}$$

Dual feasible

ⓓ $u_1, u_2 \geq 0$

Find the KKT conditions for

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & \boldsymbol{x}^T Q \boldsymbol{x} - \boldsymbol{c}^T \boldsymbol{x} \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{b} \quad (\mu) \\ & C\boldsymbol{x} \geq \boldsymbol{d} \quad (\lambda_1) \Leftrightarrow d - Cx \leq 0 \\ & \boldsymbol{x} \geq 0 \quad (\lambda_2) \Leftrightarrow -x \leq 0 \end{aligned}$$

$$L(x, \lambda, \mu) = x^T Q x - c^T x + \mu^T (Ax - b) + \lambda_1^T (d - Cx) - \lambda_2^T x$$

$$\lambda_1, \lambda_2 \geq 0$$

KKT Conditions (a) + (b) + (c) + (d)

Main:

$$\nabla_x L(x, \lambda, \mu) = 2Qx - c + A^T\mu - C^T\lambda_1 - \lambda_2$$

(a) $$2Qx - c + A^T\mu - C^T\lambda_1 - \lambda_2 = 0$$

Complementary Slackness

(b) $$\begin{cases} \lambda_1^T (d - Cx) = 0 \\ \lambda_2^T x = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \lambda_{1i} (d_i - C_i^T x_i) = 0 \\ \lambda_{2i} x_i = 0 \end{cases} \quad \forall i$$

Primal feasible

(c) $Ax = b$, $Cx \geq d$, $x \geq 0$

Dual feasible

(d) $\lambda_1, \lambda_2 \geq 0$

Example 3: Cylinder Volume

We want to build a cylinder with the maximum volume, with its surface area no larger than C.

- Decision variables: r (the radius of the base) and h (height).
- Then the optimization problem is:

$\min. -\pi r^2 h$

$$\begin{array}{lll} \text{maximize}_{r,h} & \pi r^2 h & \\ \text{subject to} & 2\pi r^2 + 2\pi rh \leq C & (\lambda_1) \\ & r, h \geq 0 & (\lambda_2, \lambda_3) \end{array}$$

$$L(r, h, \lambda) = -\pi r^2 h + \lambda_1 (2\pi r^2 + 2\pi rh - C) - \lambda_2 r - \lambda_3 h$$

$$\lambda_1, \lambda_2, \lambda_3 \geq 0$$

KKT Conditions:

Main.

$$\begin{cases} \frac{d L(r,h,\lambda)}{dr} = -2\pi r h + 4\lambda_1 \pi r + 2\lambda_1 \pi h - \lambda_2 = 0 & (1) \\ \frac{d L(r,h,\lambda)}{dh} = -\pi r^2 + 2\lambda_1 \pi r - \lambda_3 = 0 & (2) \end{cases}$$

Complementary Slackness

$$\begin{cases} \lambda_1 \cdot (2\pi r^2 + 2\pi r h - C) = 0 & (3) \leftarrow \\ \lambda_2 \cdot r = 0, \quad \lambda_3 \cdot h = 0 & (4) \end{cases}$$

Primal & dual feasible

$$2\pi r^2 + 2\pi r h \le C, \quad r \ge 0, \quad h \ge 0$$

$$\lambda_1, \lambda_2, \lambda_3 \ge 0$$

$r > 0, h > 0 \overset{(4)}{\Longrightarrow} \underline{\lambda_2 = 0, \quad \lambda_3 = 0}$ For sure

From (3), Case I: $\lambda_1 = 0$, Plug in (2)

$$-\pi r^2 = 0 \Rightarrow r = 0 \quad \text{Impossible}$$

Case II:

$$\boxed{\begin{cases} 2\pi r^2 + 2\pi r h = C & (3') \\ -2\pi r h + 4\lambda_1 \pi r + 2\lambda_1 \pi h = 0 & (1) \\ -\pi r^2 + 2\lambda_1 \pi r = 0 & (2) \end{cases}}$$

From (2) $2\lambda_1 r = r^2 \Rightarrow \underline{r = 2\lambda_1}$

Plug in (1)

$$-2\cancel{\pi}(2\cancel{\lambda_1})h + 4\lambda_1\cancel{\pi}(2\cancel{\lambda_1}) + 2\cancel{\lambda_1}\cancel{\pi}h = 0$$

$$\Rightarrow -4h + 8\lambda_1 + 2h = 0$$

$$\Rightarrow h = 4\lambda_1$$

Plug in (3')

$$2\pi(4\lambda_1^2) + 4\pi(2\lambda_1)4\lambda_1 = 0$$

$$\Rightarrow \lambda_1 = \sqrt{\frac{c}{24\pi}}$$

$$r = 2\lambda_1 = 2\sqrt{\frac{c}{24\pi}}$$

$$h = 4\lambda_1 = 4\sqrt{\frac{c}{24\pi}}$$

optimal

Example 4: Power Allocation

We have a collection of n communication channels and we need to decide how much power to allocate to each of them

- The capacity (communication rate) of channel i is $\log(\alpha_i + x_i)$ with a given $\alpha_i > 0$ and when x_i is allocated to it, and we have a budget constraint $\boldsymbol{e}^T \boldsymbol{x} = 1$, $\boldsymbol{x} \geq 0$

The optimization problem is:

$\to \min -\sum_i \log(\alpha_i + x_i)$

$$\begin{array}{rll} \text{maximize}_{\boldsymbol{x}} & \sum_{i=1}^n \log(\alpha_i + x_i) & \\ \text{subject to} & \sum_{i=1}^n x_i = 1 & (\mu) \to \text{scalar} \\ & \boldsymbol{x} \geq 0 & (\lambda) \to \text{vector} \end{array}$$

$$L(x, \lambda, \mu) = -\sum_{i=1}^n \log(\alpha_i + x_i) - \lambda^T x + \mu\left(\sum_{i=1}^n x_i - 1\right)$$

$\lambda \geq 0$

$$L(x,\lambda,\mu) = -\sum_i \log(\alpha_i + x_i) - \sum_i \lambda_i x_i + \mu\left(\sum_{i=1}^{n} x_i - 1\right)$$

KKT Conditions (a)+(b)+(c)

Main: $\nabla_x L(x,\lambda,\mu) = \begin{bmatrix} \frac{\partial L}{\partial x_1} \\ \vdots \\ \frac{\partial L}{\partial x_n} \end{bmatrix}$

$$= \begin{bmatrix} -\frac{1}{\alpha_1+x_1} - \lambda_1 + \mu \\ \vdots \\ -\frac{1}{\alpha_n + x_n} - \lambda_n + \mu \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$\Rightarrow$ (a) $\underbrace{-\frac{1}{\alpha_i + x_i} - \lambda_i + \mu = 0,}$ $\forall i = 1, \cdots, n$

Complementary Slackness:

(b) $\lambda_i x_i = 0, \quad \forall i$ $\begin{cases} \lambda_i = 0 & ① \\ x_i = 0 & ② \end{cases}$

Primal & dual feasibility

(c) $\sum_i x_i = 1, \quad x \geq 0, \quad \lambda \geq 0$

① If $\lambda_i = 0$

$$(a) \Rightarrow -\frac{1}{\alpha_i + x_i} + \mu = 0$$

$$x_i = \frac{1}{\mu} - \alpha_i, \quad \forall i$$

require: $x_i \geq 0 \Rightarrow \frac{1}{\mu} - \alpha_i \geq 0 \Rightarrow \mu \leq \frac{1}{\alpha_i} \quad \forall i$

② If $x_i = 0$

$$(a) \quad -\frac{1}{\alpha_i} - \lambda_i + \mu = 0$$

$$\lambda_i = \mu - \frac{1}{\alpha_i}$$

require $\lambda_i \geq 0 \Rightarrow \mu - \frac{1}{\alpha_i} \geq 0 \Rightarrow \mu \geq \frac{1}{\alpha_i} \quad \forall i$

$$x_i^* = \begin{cases} 0 & \text{if } \mu \geq \frac{1}{\alpha_i} \\ \frac{1}{\mu} - \alpha_i & \text{if } \mu < \frac{1}{\alpha_i} \end{cases}$$

$$= \max\left\{0, \frac{1}{\mu} - \alpha_i\right\}$$

Use $\sum_i x_i^* = \sum_i \max\left\{0, \frac{1}{\mu} - \alpha_i\right\} = 1$

to find out μ.

Outline

Last Topic: Algorithms

Discuss how to solve nonlinear optimization problems.

- We have shown that in many cases, KKT conditions can be used to solve the optimization problem
- However, those are ad hoc situations. In most cases, we cannot directly find the optimal solution from the KKT conditions
- We want to have a robust procedure (an algorithm) that guarantees to solve the optimization problem.

Unconstrained Problems

We start with the unconstrained problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x})$$

We are going to study the following methods:

- Bisection search
- Golden section search
- Gradient descent method
- Newton's method
- More (time permitted)

General Solution Idea

Typically, optimization algorithms are *iterative* procedures.

- Start from some feasible point $\boldsymbol{x}_0$, then generate a sequence of $\{\boldsymbol{x}_k\}$
- The sequence terminates when either no progress can be made or when we know that the current solution is already satisfactory
- Typically, we want to have $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$, i.e., each step we can improve the objective value.
- And hopefully, we want the sequence $\{\boldsymbol{x}_k\}$ to *converge* to a local minimizer $\boldsymbol{x}^*$ (or global minimizer).

Outline

Single Variable Problem

Assume $f(x)$ is a single variable function.

Objective: find a local minimizer of $f(x)$.

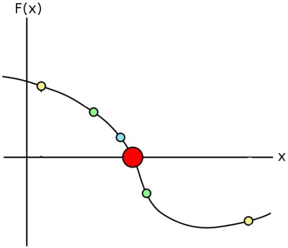
We introduce two methods:

- Bisection method
- Golden section method

Bisection Method

Bisection method uses the idea that the local minimizer must satisfy the FONC: $f'(x) = 0$.

Therefore, the problem becomes a root-finding problem for $g(x) = f'(x)$.



Root Finding Algorithm: Bisection Method

Assume one can find x_ℓ and x_r such that $g(x_\ell) < 0$ and $g(x_r) > 0$. By intermediate value theorem, if $g(\cdot)$ is continuous, there must exist a root of $g(\cdot)$ in $[x_\ell, x_r]$.

Bisection method:

1. Define $x_m = \frac{x_\ell + x_r}{2}$
2. If $g(x_m) = 0$, then output x_m
3. Otherwise
    - If $g(x_m) > 0$, then let $x_r = x_m$
    - If $g(x_m) < 0$, then let $x_\ell = x_m$
4. If $|x_r - x_\ell| < \epsilon$. stop and output $\frac{x_\ell + x_r}{2}$, otherwise go back to Step 1

One can also set the stop criterion based on $|g(x)| \leq \varepsilon$

Bisection Method

In the bisection method, each iteration will divide the search interval to half.

Therefore, to find an ϵ approximation of x^*, we need at most $\log_2 \frac{x_r - x_\ell}{\epsilon}$ iterations

Applying bisection method to $g = f'$, one can find a critical point satisfying FONC (approximately).

- If f is convex, we can find the global minimizer of $f(x)$ (approximately).
- Although simple, the bisection method is very useful in practice because it is easy to implement.

Golden Section Method

One drawback of using the bisection method to solve (single variable, unconstrained) optimization problems is that it requires the knowledge (and computation) of $f'(x)$.

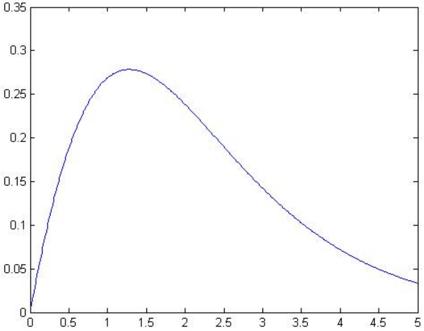
- Sometimes, we don't have $f'(x)$ available. For example, $f(x)$ sometimes is only a *black box*, which does not admit an analytical form (thus the derivative is hard to compute).

However, if we know that $f(x)$ has a unique local optimal x^* in the range $[x_\ell, x_r]$, then we still have very efficient way to find it out.

- We call such f *unimodal* on $[x_\ell, x_r]$
- Unimodal function has the property that the local optimal is global optimal (convex function is always unimodal).

Example of Unimodal Function (Maximization)

Consider $f(x) = \frac{xe^{-x}}{1+e^{-x}}$:



This is a unimodal function, but not a concave function.

Golden Section Method

Assume we start with $[x_\ell, x_r]$. Assume $0 < \phi < 0.5$.

1. Set $x'_\ell = \phi x_r + (1-\phi)x_\ell$ and $x'_r = (1-\phi)x_r + \phi x_\ell$.
2. If $f(x'_\ell) < f(x'_r)$, then the minimizer must lie in $[x_\ell, x'_r]$, so set $x_r = x'_r$
3. Otherwise, the minimizer must lie in $[x'_\ell, x_r]$, so set $x_\ell = x'_\ell$.
4. If $x_r - x_\ell < \epsilon$, output $\frac{x_\ell + x_r}{2}$, otherwise go back to Step 1.

If we want to reuse the computation (i.e.,$x''_r = x'_\ell$), then we set $\phi = \frac{3-\sqrt{5}}{2}$, and $1 - \phi = \frac{\sqrt{5}-1}{2} = 0.618$ (where the name comes from).

Higher Dimensional Problems

Next we consider the high-dimensional problem

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$$

- There is not a clear bisection or golden section in that case

Solution idea:

- Each time, we first find a search direction.
- Then we search for a good solution along that direction.

General Framework for High Dimensional Search

From $\boldsymbol{x}^0$, we generate a sequence of points:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k \boldsymbol{d}^k.$$

We call $\boldsymbol{d}^k$ the search direction (a vector) and α_k the step size (a positive scalar).

- The key is to choose proper $\boldsymbol{d}^k$ at each iteration, such that $f(\boldsymbol{x}^{k+1}) < f(\boldsymbol{x}^k)$.
- $\boldsymbol{d}^k$ typically depends on $\boldsymbol{x}^k$
- Then α_k may be chosen in accordance with some line (one-dimension) search rules.

Our goal is to find the proper $\boldsymbol{d}^k$ and α_k at each iteration so that the sequence $\{x^k\}$ converges to a local minimizer x^* (or global minimizer or something meaningful).

Some Useful Concepts: Convergent Sequences

Definition

Let $\{\boldsymbol{x}_k\}$ be a sequence of real vectors. Then $\{\boldsymbol{x}_k\}$ converges to $\boldsymbol{x}^*$ if and only if for all real numbers $\epsilon > 0$, there exists a positive integer K such that $||\boldsymbol{x}_k - \boldsymbol{x}^*|| < \epsilon$ for all $k \geq K$.

In all our discussions, we assume $||\boldsymbol{x}||$ is the 2-norm of $\boldsymbol{x} = (x_1, ..., x_n)$, which means:

$$||\boldsymbol{x}|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Example of convergence:

- $x_k = 1/k \to 0$
- $x_k = (1/2)^k \to 0$