# Introduction to Optimization

## Week5 - Nonlinear Programs, Optimality conditions

Sung-Pil Hong

Management Science/Optimization Lab
IE department
Seoul National University

© 2022 Sung-Pil Hong. All rights reserved.

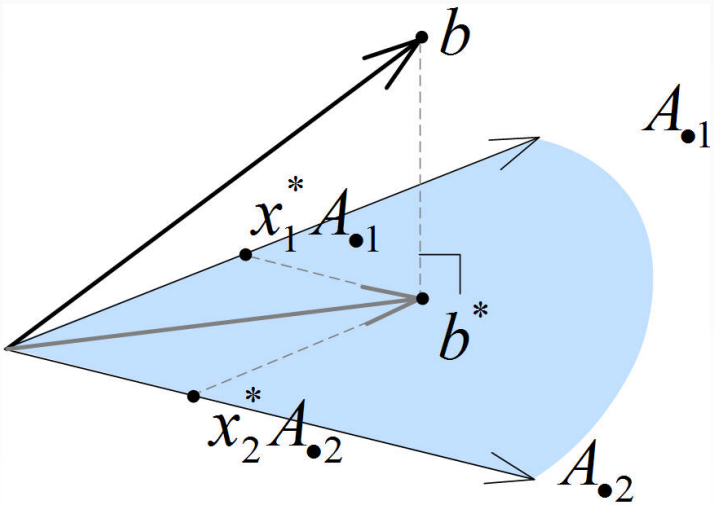
# Principles of nonlinear optimization

© 2022 Sung-Pil Hong. All rights reserved.

Suppose a linear system $Ax = b$ ($A \in \mathbb{R}^{m \times n}$) has no solution. Commonly we then settle for the value of $x$ that minimizes the error defined as the distance between the right and left hand vectors.

$$\min \|Ax - b\|_2. \tag{1.1}$$

Problem (1.1), known as the *least square problem* is the problem of computing the point $b^*$ in the column space $C(A)$ of $A$ that is closest to $b$. The point $b^*$ is the base of the perpendicular line from $C(A)$ to $b$, also called the *projection* of $b$ onto the column space. The optimal solution $(x_1^*, \ldots, x_n^*)$ is the vector of coefficients used to combine $A_{\cdot 1}, \ldots, A_{\cdot n}$ into $b^*$.




© 2022 Sung-Pil Hong. All rights reserved.

# Optimization problems

An optimization in general can be written in one of the following forms.

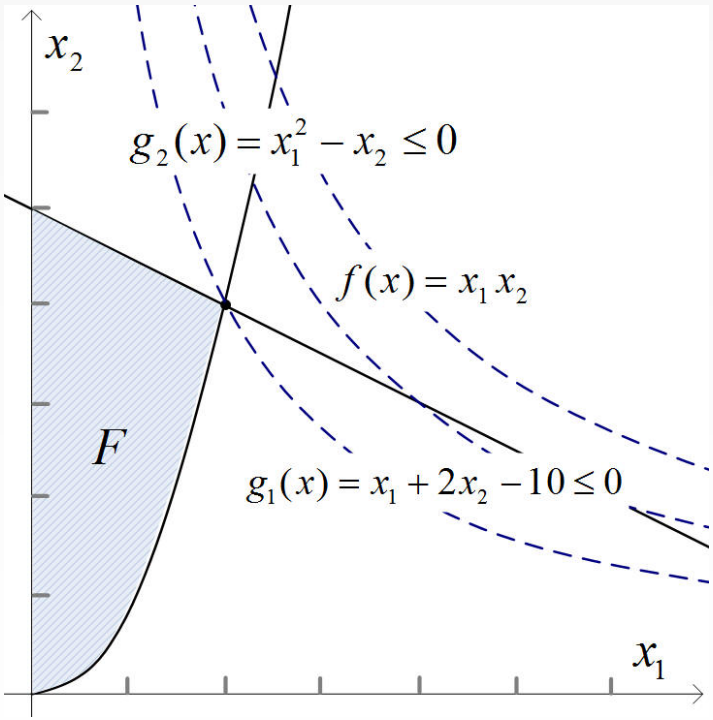
$$\begin{array}{rll} \max \text{ or } \min & f(x) & \\ \text{s.t.} & g_i(x) \leq \text{or} \geq 0, & i = 1, \ldots, m, \\ & h_j(x) = 0, & j = 1, \ldots, p. \end{array} \tag{1.2}$$

- By a feasible solution, we mean a solution $x$ satisfying every constraint. We denote by $F$ the set of feasible solutions.
- A feasible solution $x^*$ is said to be optimal if it provides the largest (or the smallest) of objective function $f(x)$.
- A feasible solution $x$ is called a local optimum if it is optimal in a neighborhood: there is a ball $B_\epsilon(x)$ centered at $x$ with radius $\epsilon > 0$ such that $f(x)$ is the smallest over $B_\epsilon(x) \cap F$.

© 2022 Sung-Pil Hong. All rights reserved.

If an optimization problem has a few variables, we can illustrate it in the space of variables using the level sets of objective function.

$$\begin{array}{rlrll} \max & f(x) = & & x_1 x_2 & \\ \text{sub. to} & g_1(x) = & x_1 & +2x_2 - 10 & \leq 0 \\ & g_2(x) = & x_1^2 & -x_2 & \leq 0 \\ & & & x \geq 0 & \end{array}$$




© 2022 Sung-Pil Hong. All rights reserved.

For a function $f : \mathbb{R}^3 \to \mathbb{R}^2$, $x = [x_1, x_2, x_3]^T \mapsto f(x) = [f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3)]^T$, its derivative $Df(\bar{x})$ is defined as a linear transformation which approximates $f$ around $x = \bar{x}$ with an error dominated by the distance from $\bar{x}$: $\|f(\bar{x} + y) - f(\bar{x}) - Df(\bar{x})y\|_2 = o(\|y\|_2)$. Specifically, $Df(\bar{x})$ is given by

$$Df(\bar{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \frac{\partial f_1}{\partial x_2}(\bar{x}) & \frac{\partial f_1}{\partial x_3}(\bar{x}) \\ \frac{\partial f_2}{\partial x_1}(\bar{x}) & \frac{\partial f_2}{\partial x_2}(\bar{x}) & \frac{\partial f_2}{\partial x_3}(\bar{x}) \end{bmatrix} \tag{1.3}$$

Definition: $h(\lambda)$ is $o(\lambda) \iff \lim_{\lambda \to 0} \frac{h(\lambda)}{\lambda} = 0$.

© 2022 Sung-Pil Hong. All rights reserved.

Gradient and hessian are important resources in designing optimization algorithms. We now take a look at some preliminary facts about gradient in an informal manner.

### Definition 1.1

**Gradient:** For $f : \mathbb{R}^n \to \mathbb{R}$, its gradient vector at $x = \bar{x}$ is

$$\nabla f(\bar{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\bar{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\bar{x}) \end{bmatrix}. \tag{1.4}$$

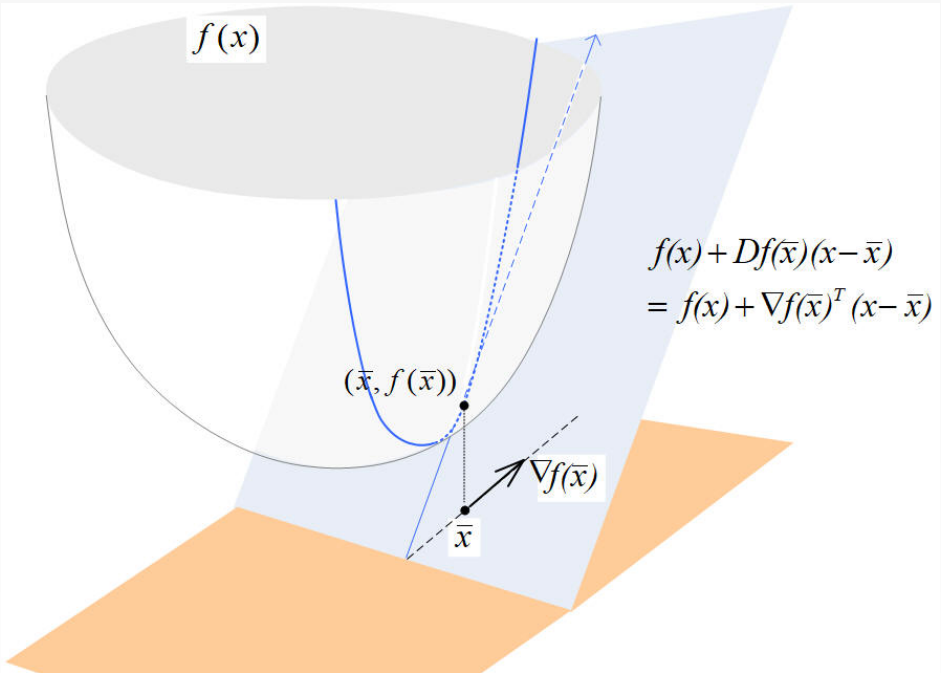
**Hessian:** For $f : \mathbb{R}^n \to \mathbb{R}$, its hessian at $x = \bar{x}$ is defined by the derivative at $x = \bar{x}$ of the gradient $x \mapsto \nabla f(x)$:

$$\nabla^2 f(\bar{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\bar{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\bar{x}) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\bar{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\bar{x}) \end{bmatrix} \tag{1.5}$$

© 2022 Sung-Pil Hong. All rights reserved.

- When $f$ is real-valued, $Df(\bar{x})$ approximates $f$ as follows:

$$f(\bar{x}+y) \approx f(\bar{x}) + \left[\frac{\partial f}{\partial x_1}(\bar{x}), \frac{\partial f}{\partial x_2}(\bar{x}), \frac{\partial f}{\partial x_3}(\bar{x})\right] y.$$




- If $f$ is a linear function $Ax$, its derivative is $A$ at every point. Thus the derivative of $c^T x$ is $c^T$.

© 2022 Sung-Pil Hong. All rights reserved.

## Proposition 1.2

*If $h : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ have derivatives, their composition $f := g \circ h : \mathbb{R}^n \to \mathbb{R}^p$, $x \mapsto g(h(x))$ also has a derivative, given by*

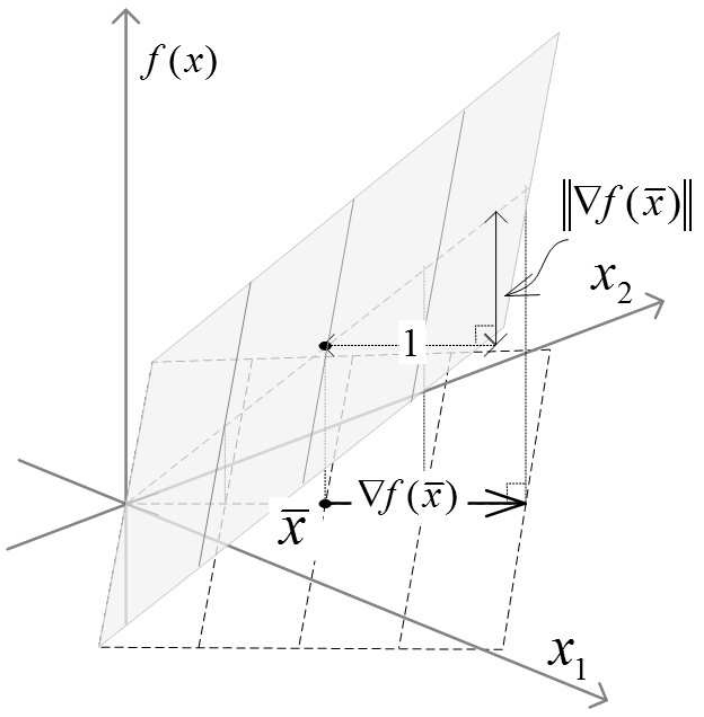
$$Df(x) = D(g \circ h)(x) = Dg(h(x))Dh(x).$$

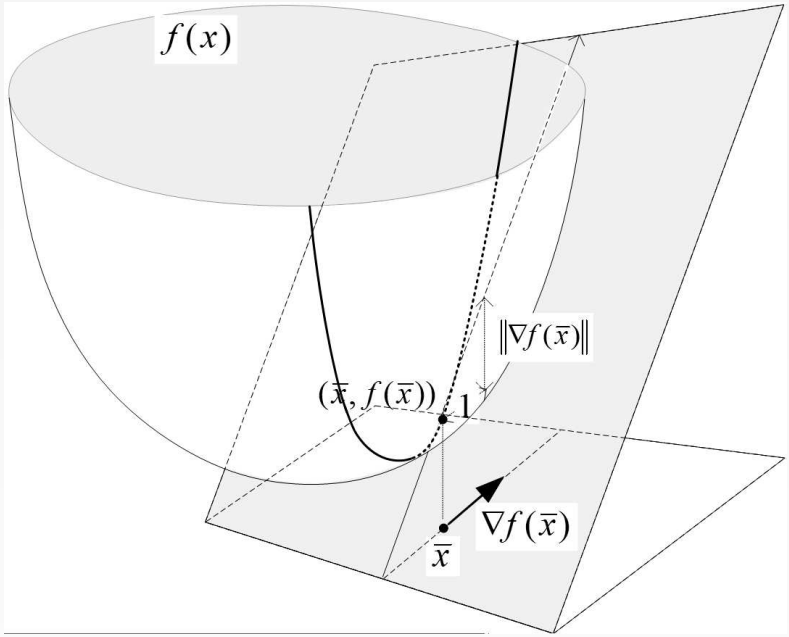
$f : \mathbb{R}^3 \to \mathbb{R}$, $Df(x) = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}] \in \mathbb{R}^{1\times 3}$. $g : \mathbb{R} \to \mathbb{R}^3$, $Dg(t) = [g_1'(t), g_2'(t), g_3'(t)]^T \in \mathbb{R}^{3\times 1}$. Then the function $t \mapsto (f \circ g)(t) = f(g_1(t), g_2(t), g_3(t))$ has derivative

$$\begin{aligned} h'(t) &= Df(g(t))Dg(t) \\ &= [\tfrac{\partial f}{\partial x_1}(g(t)), \tfrac{\partial f}{\partial x_2}(g(t)), \tfrac{\partial f}{\partial x_3}(g(t))] \begin{bmatrix} g_1'(t) \\ g_2'(t) \\ g_3'(t) \end{bmatrix} \\ &= \nabla f(g(t))^T Dg(t). \end{aligned}$$

If $g(t) = x + ty$ ($x, y \in \mathbb{R}^3$), since $Dg(t) = y$ we have $h'(t) = \nabla f(x + ty)^T y$. We call $h'(0) = \nabla f(x)^T y$, the directional derivative of $f$ at $x$ into $y$.

© 2022 Sung-Pil Hong. All rights reserved.



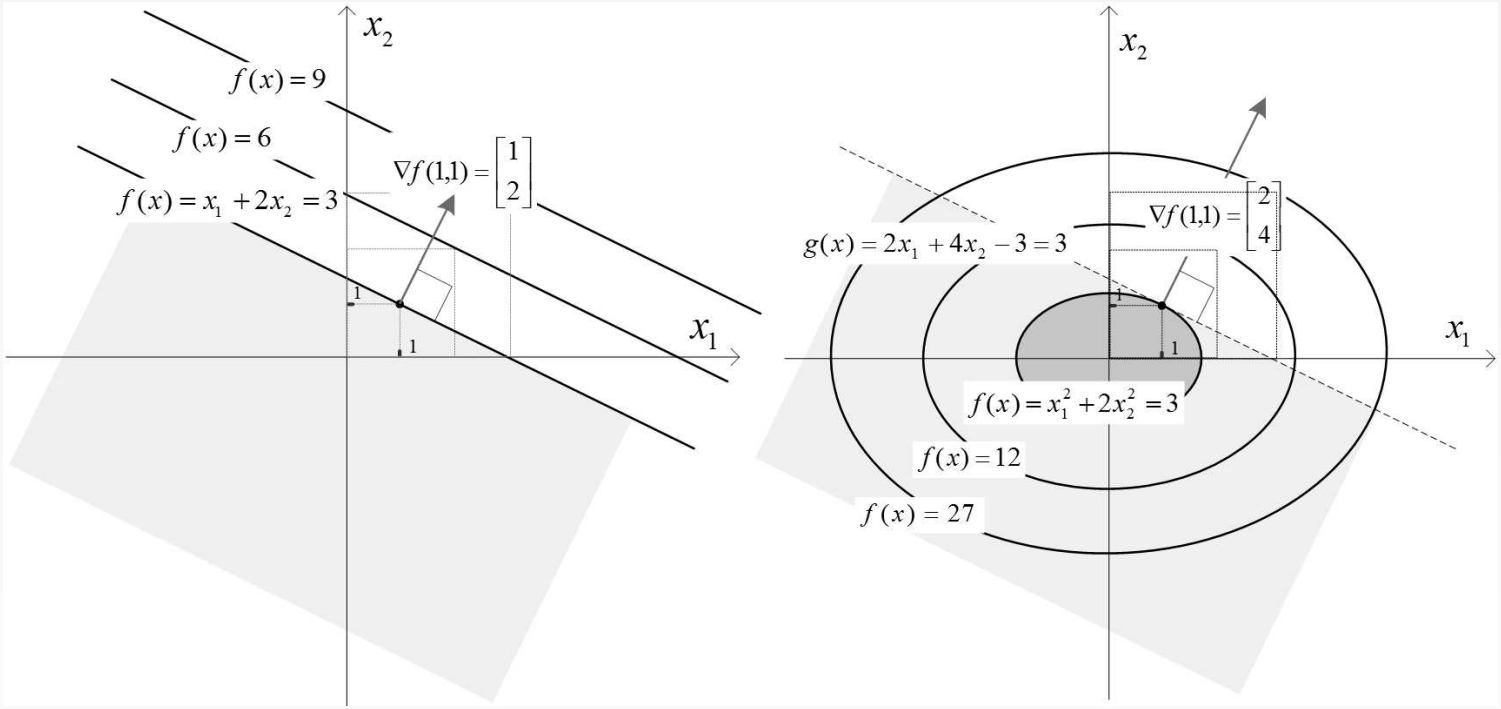




The gradient $\nabla f(x) = [1, 1]^T$ of a linear function $f(x_1, x_2) = x_1 + x_2$ is the direction of the fastest growth and is normal to the contour at every point. The growth rate is $\|\nabla f(1, 1)\|_2 = \sqrt{2}$.

In general, the gradient $\nabla f(\bar{x})$ of a function $f(x)$ at $x = \bar{x}$, which is identical to the gradient of the linear function whose graph is tangent plane to the graph of $f(x)$ at $(\bar{x}, f(\bar{x}))$, is the direction $f$ of the largest instantaneous rate of growth at $x = \bar{x}$. The rate is $\|\nabla f(\bar{x})\|_2$.

© 2022 Sung-Pil Hong. All rights reserved.

Level sets of $f(x) = x_1 + 2x_2$ and $f(x) = x_1^2 + 2x_2^2$.




A linear function $f(x) = c^T x$ increases in the direction $y$ at the rate of $c^T y / \|y\|$ from any point. Hence if $\nabla f(\bar{x})^T y < 0$, $f$ decreases at a constant rate along the half line from $\bar{x}$ in the direction $y$. In general, when $f$ is nonlinear, although not decreasing at a constant rate, there is an interval immediately after $\bar{x}$ along the line on which $f$ is smaller than $f(\bar{x})$.

© 2022 Sung-Pil Hong. All rights reserved.

## Proposition 1.3

*If* $\nabla f(\bar{x})^T y < 0$, $y$ *is a descent direction from* $x = \bar{x}$: $\exists\ \bar{\lambda} > 0\ :\ f(\bar{x} + \lambda y) < f(\bar{x})\ \forall\ 0 < \lambda < \bar{\lambda}$.

**Proof:** We take the single-variable case as given. For the $x \in \mathbb{R}^n$ case, consider $g(\lambda) := f(\bar{x} + \lambda y)$, a function of $\lambda \in \mathbb{R}$. By the chain rule,

$$g'(0) = \nabla f(\bar{x})^T y < 0. \tag{1.6}$$

By the proposition for the single-variable case, there is $\bar{\lambda} > 0$ such that

$$\forall\ 0 < \lambda \leq \bar{\lambda},\ f(\bar{x} + \lambda y) < f(\bar{x}).\ \square$$

## Exercise 1.4

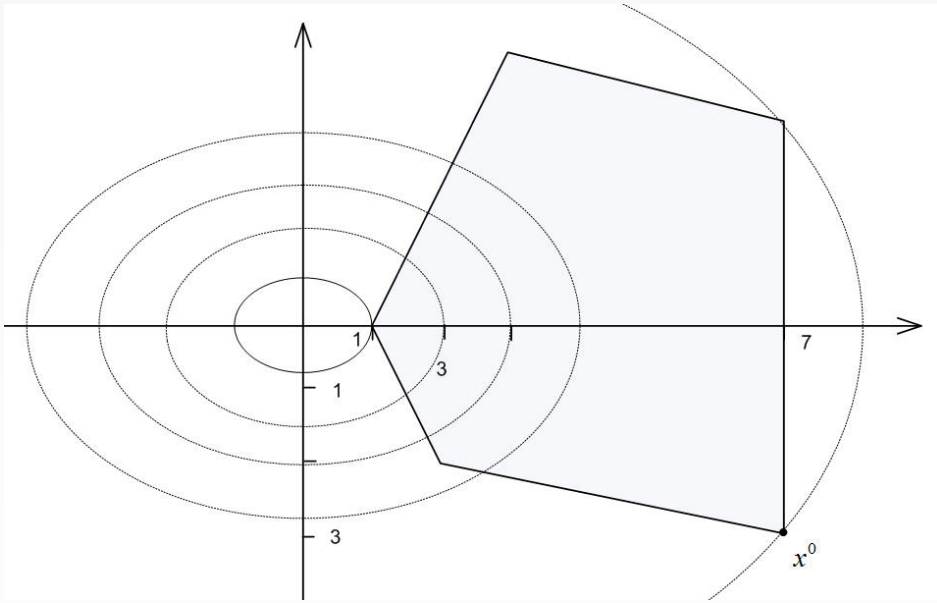
1. *Restate the proposition for ascent directions and provide a proof.*
2. *Sketch the gradient, contours, and the ascent directions of* $f(x) = (x_1 - 2x_2)^2$ *at* $x = (1, 1)$.

© 2022 Sung-Pil Hong. All rights reserved.

## Exercise 1.5

*Indicate the descent directions of the objective function from* $x^0 = [7, -3]^T$.

$$
\begin{array}{lllll}
\min & \frac{3}{4}x_1^2 & +x_2^2 & & \\
\text{sub.to} & 2x_1 & -x_2 & \geq 2, & \\
& 2x_1 & +x_2 & \geq 2, & \\
& x_1 & +4x_2 & \leq 19, & \\
& x_1 & & \leq 7, & \\
& x_1 & +5x_2 & \geq -8. &
\end{array}
$$

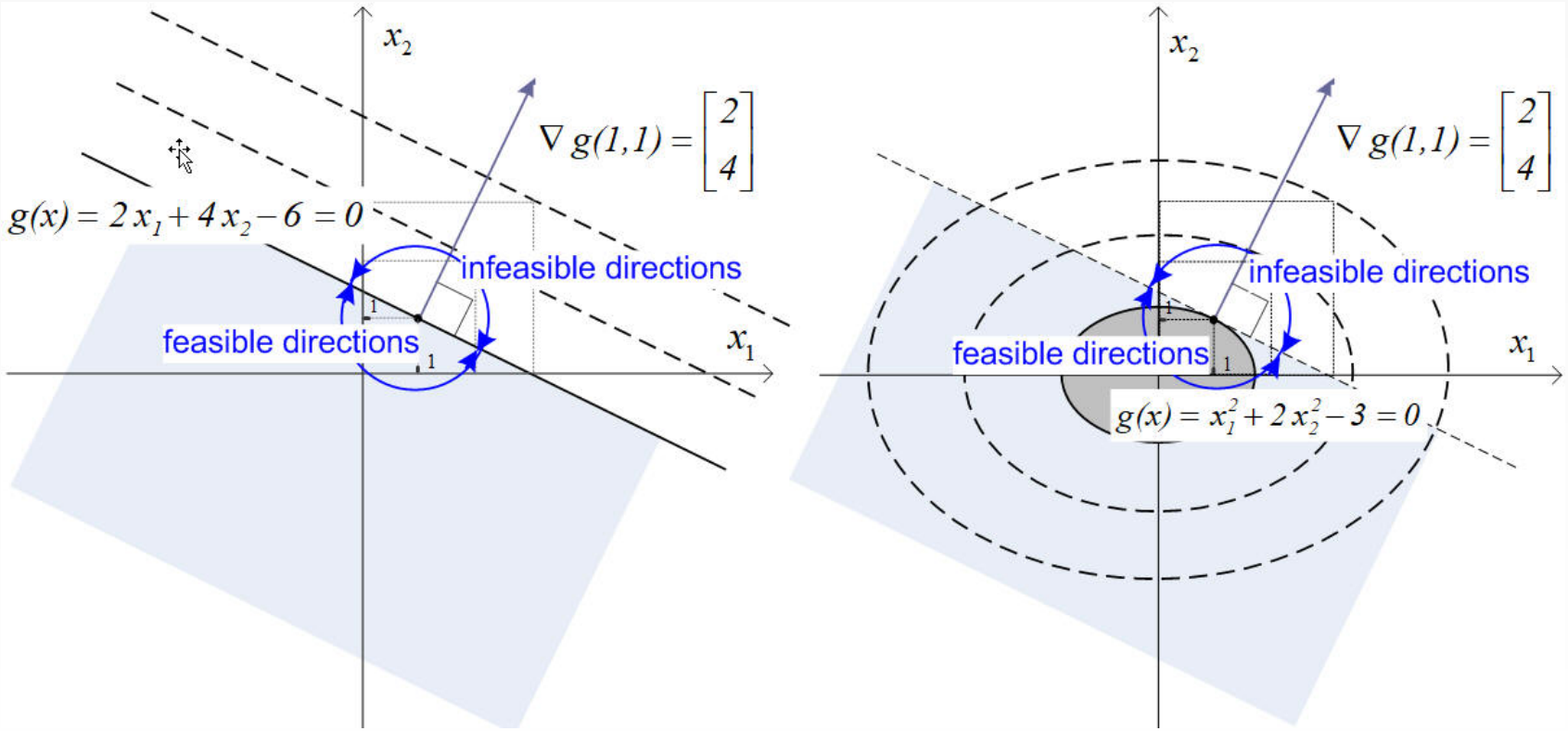



© 2022 Sung-Pil Hong. All rights reserved.

## Definition 1.6

We call $y$ a *feasible direction* from $x \in F$ if we can move in the direction $y$ for a positive distance maintaining feasibility; that is, $\exists\ \bar{\lambda} > 0$: $x + \lambda y \in F$, $\forall\ 0 < \lambda < \bar{\lambda}$.

According to Proposition 1.3, if $g(x) \leq 0$ $(g(x) \geq 0)$ is an active constraint of $\bar{x}$, any $y$ such that $\nabla g(\bar{x})^T y < 0$ $(\nabla g(\bar{x})^T y > 0$, resp.) is a feasible direction of $\bar{x}$.

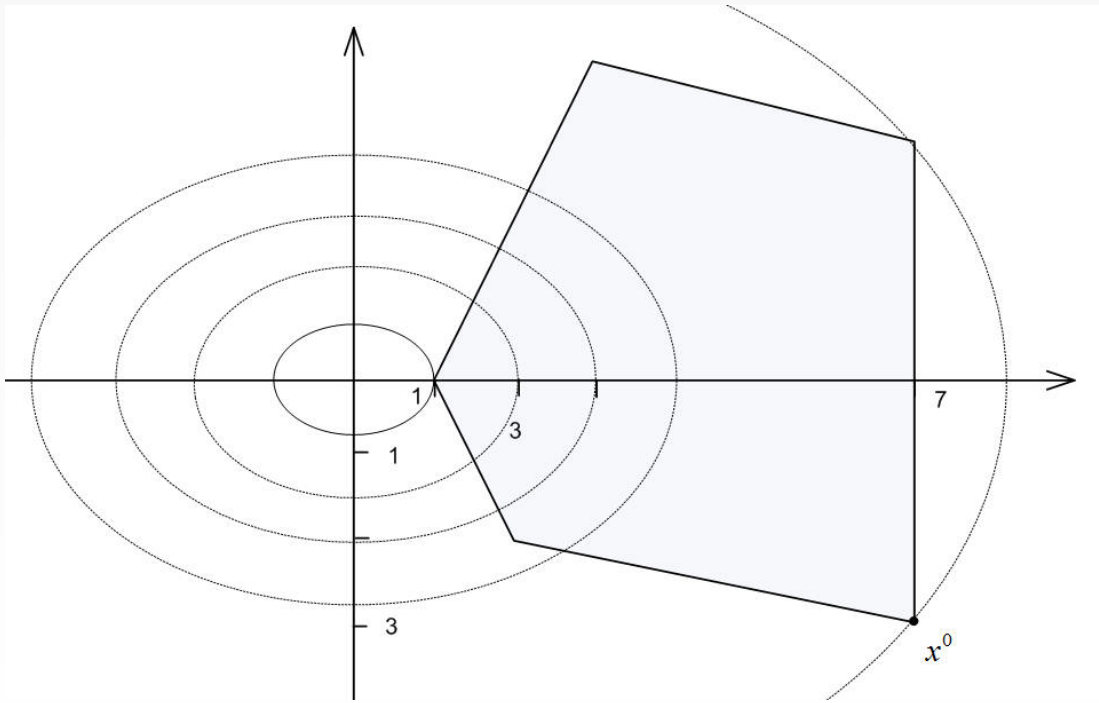



© 2022 Sung-Pil Hong. All rights reserved.

If there is more than one active constraint $g_i(x) \leq 0$, then a direction $y$ satisfying $\nabla g_i^T(\bar{x})y < 0$ for *all* $i$ is a feasible direction of $\bar{x}$. The inactive constraints do not restrict feasible direction set, as $\bar{x}$ satisfies each of them with a strict inequality.

**Exercise 1.7**

*Compute the feasible directions of $x^0$ in the optimization problem in Exercise 1.5. Is $x^0$ optimal? Explain.*




© 2022 Sung-Pil Hong. All rights reserved.

# Optimality conditions

© 2022 Sung-Pil Hong. All rights reserved.

There is no known necessary and sufficient optimality condition for a general optimization problem which we can recognize or implement in an efficient manner. And it is believed that such a condition is unlikely to exist.

Today, we are interested in a necessary condition of optimality, known as *KKT optimality conditions*. We will derive it by combining the previous observations and, interestingly, the duality theory of linear programming.

To this end, we define improving directions.

**Definition 2.1**

For minimization problems, *improving directions* = descent directions $\cap$ feasible directions. For maximization problems, ....

© 2022 Sung-Pil Hong. All rights reserved.

As a preliminary, we first show that any linear inequality system defined by linearly independent vectors has an interior feasible solution:

### Proposition 2.2

*If* $v_1, \ldots, v_m$ *are linearly independent,* $K^\circ = \{x : v_1^T x < 0, \ldots, v_m^T x < 0\} \neq \emptyset$.

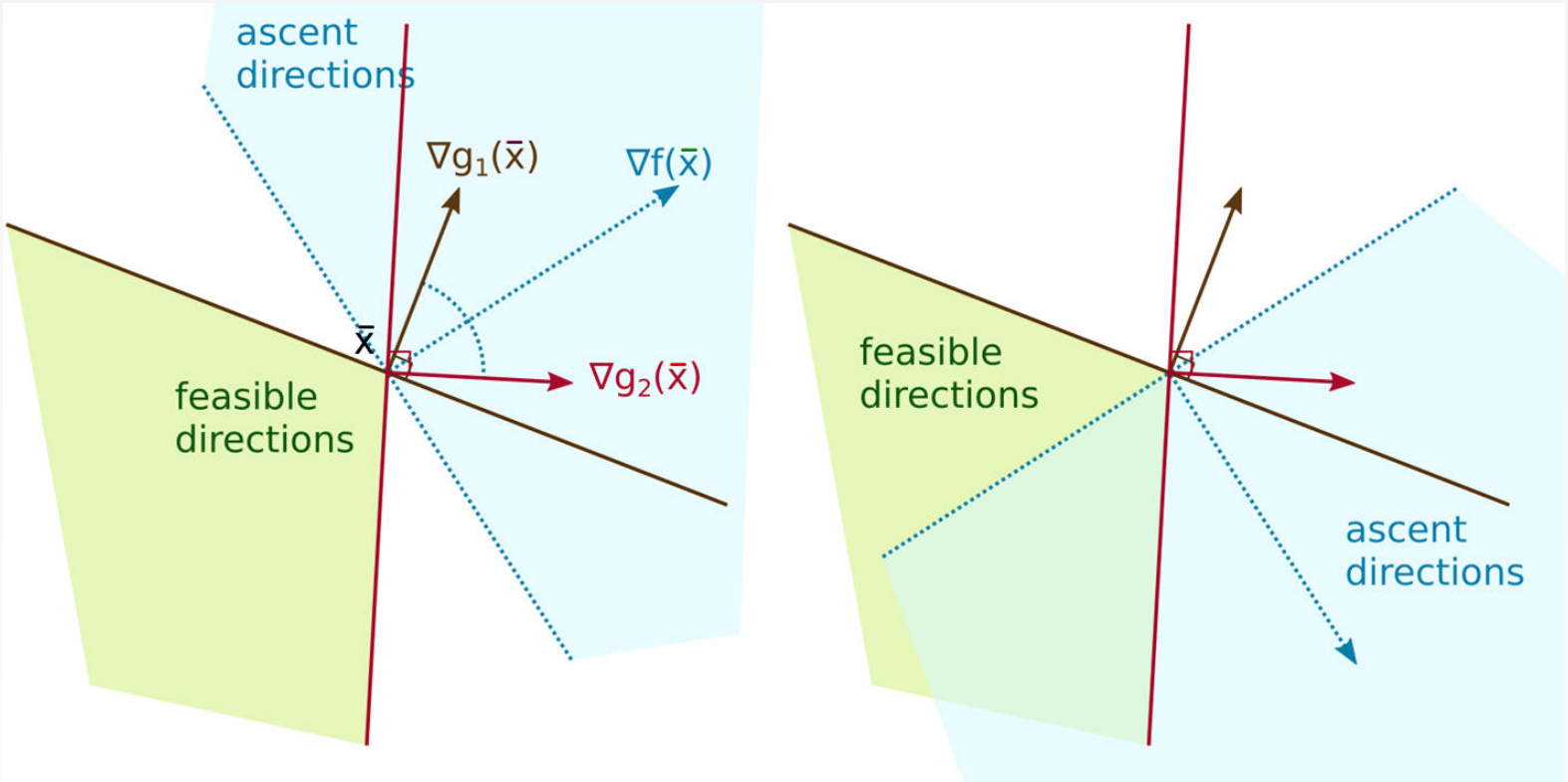
**Proof**: Assume, to get a contradiction, that $K^\circ = \emptyset$. Take any $1 \leq i' \leq m$. Then $\forall\ x$: $V_{-i'}^T x < 0$ we have $v_{i'}^T x \geq 0$, where $V_{-i'} = \left[v_1, \ldots, v_{i'-1}, v_{i'+1}, \ldots, v_m\right]$. Since linear functions are continuous, $0 = \min\{v_{i'}^T x : V_{-i'}^T x \leq 0\}$. The dual problem is $\max\ \{0^T y : V_{-i'} y = v_{i'}, y \leq 0\}$. Hence, by strong duality, there is $y \leq 0$: $V_{-i'} y = v_{i'}$. This contradicts the independence of $v_i$'s. $\square$

© 2022 Sung-Pil Hong. All rights reserved.

The KKT conditions are about nonexistence of a particular type of improving direction at local optima. To illustrate the idea, we consider a problem $\max \{f(x) : g_1(x) \leq 0, \ldots, g_m(x) \leq 0\}$ and its feasible solution $\bar{x}$ having the first two constrains as the active constraints.

The vectors having a negative inner product with each gradient of the active constraints of $\bar{x}$ are feasible directions. If the gradients are linearly independent, by Proposition 2.2, their set is nonempty. Thus if $\bar{x}$ is a (local) optimum, there is no such a feasible direction whose inner product with the gradient of objective function is positive and which is hence an ascent direction.

© 2022 Sung-Pil Hong. All rights reserved.




The green region represents the feasible directions $x$ whose inner product with the gradient of every active constraint at $x = \bar{x}$ is negative: $\nabla g_1(\bar{x})^T x < 0$ and $\nabla g_2(\bar{x})^T x < 0$. The blue region indicate the directions $x$ whose inner product with the gradient of objective function at $x = \bar{x}$ is positive, $\nabla f(\bar{x})^T x$, hence ascent directions.

© 2022 Sung-Pil Hong. All rights reserved.

As in the picture, if $\bar{x}$ is a local optimum, the blue and red regions should not intersect. That is $\nabla f(\bar{x})^T x \leq 0$ for every $x$: $\nabla g_1(\bar{x})^T x < 0$, and $\nabla g_2(\bar{x})^T x < 0$. Thus from the continuity of linear function and the feasibility of $0$, we conclude $0 = \max \{\nabla f(\bar{x})^T x : \nabla g_1(\bar{x})^T x \leq 0, \nabla g_2(\bar{x})^T x \leq 0\}$.

The dual problem is $\min \Big\{ 0^T \lambda : [\nabla g_1(\bar{x}) \nabla g_2(\bar{x})]\, \lambda = \nabla f(\bar{x}),\ \lambda_1 \geq 0,\ \lambda_2 \geq 0 \Big\}$. The strong duality implies the existence of $\lambda_1$ and $\lambda_2$ such that $\nabla f(\bar{x}) = \nabla g_1(\bar{x})\lambda_1 + \nabla g_2(\bar{x})\lambda_2$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$. This condition is equivalent to the existence of $\lambda = (\lambda_1, \ldots. \lambda_m)$ such that

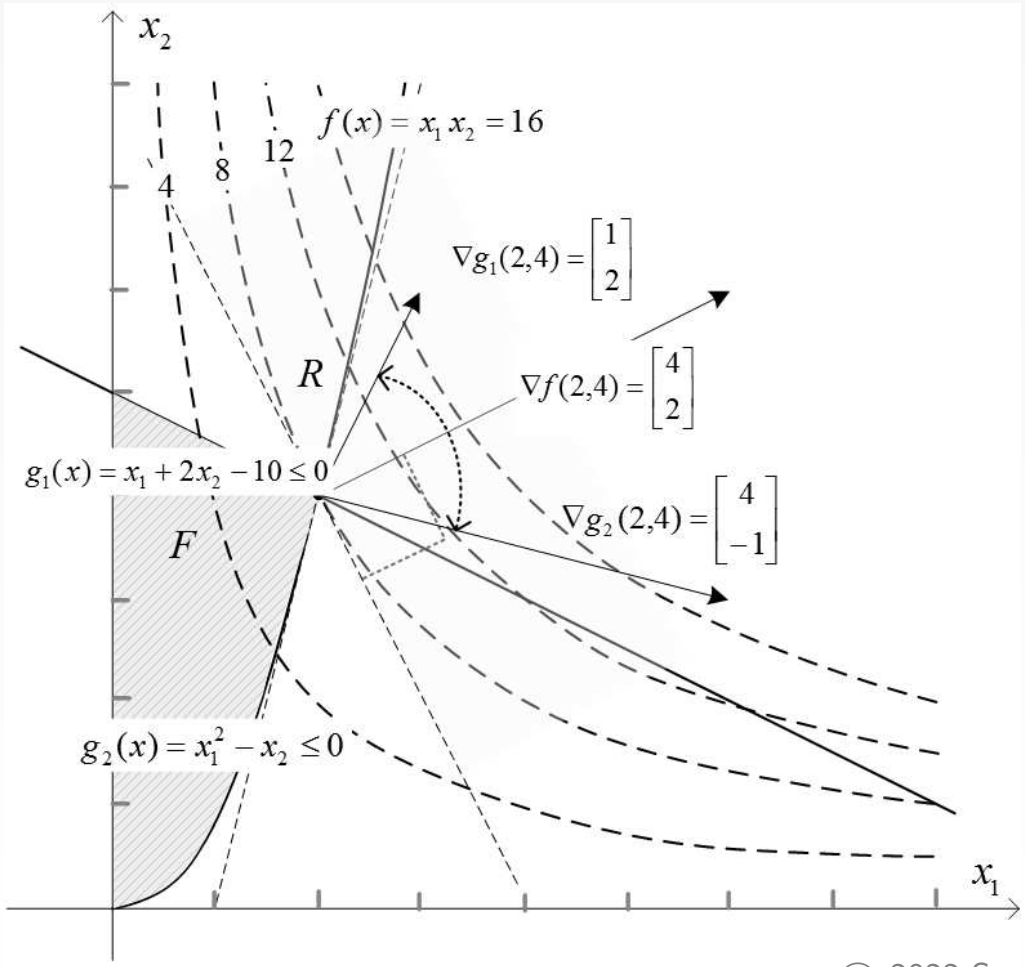
$$\begin{aligned} \nabla f(\bar{x}) &= \nabla g_1(\bar{x})\lambda_1 + \cdots \nabla g_m(\bar{x})\lambda_m, \\ \lambda &\geq 0, \text{ and} \\ g(\bar{x})^T \lambda &= 0. \end{aligned}$$

Clearly, the arguments can be applied to a general case. And we have established the KKT conditions for inequality constrained optimization.

© 2022 Sung-Pil Hong. All rights reserved.

Let's review the arguments through the next nonlinear optimization.

$$\begin{array}{lllll} \max & f(x) = & x_1 x_2 & & \\ \text{sub. to} & g_1(x) = & x_1 + 2x_2 - 10 & \leq 0 \\ & g_2(x) = & x_1^2 - x_2 & \leq 0 \\ & & x \geq 0 \end{array} \tag{2.7}$$




© 2022 Sung-Pil Hong. All rights reserved.

First note that the gradients $[1, 2]$ and $[4, -1]$ of the active constraints of $\bar{x} = (2, 4)$ are linearly independent. Hence we have a nonempty set of directions having a negative inner product with them. Thus since $\bar{x} = (2, 4)$ is a (local) optimum, there is no such a feasible direction whose inner product with the objective gradient $[4, 2]^T$ is positive and which is hence an ascent direction at $\bar{x} = (2, 4)$.

Hence we have $\nabla f(\bar{x})^T x = [2, 4]x \leq 0$ for every $x$ such that $\nabla g_1(\bar{x})^T x = [1, 2]x < 0$, $\nabla g_2(\bar{x})^T x = [4, -1]x < 0$. Since $\nabla f(\bar{x})^T x$ is a continuous function and $x = 0$ is a feasible solution, it implies $0 = \max\{[2, 4]x : [1, 2]x \leq 0,\ [4, -1]x \leq 0\}$.

The dual problem is $\min\left\{0^T\lambda : \begin{bmatrix} 1 & 4 \\ 2 & -1 \end{bmatrix}\lambda = \begin{bmatrix} 2 \\ 4 \end{bmatrix},\ \lambda \geq 0\right\}$. The strong duality implies the existence of $y$:

$$\begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\lambda_1 + \begin{bmatrix} 4 \\ -1 \end{bmatrix}\lambda_2,\ \ \lambda_1 \geq 0,\ \lambda_2 \geq 0.$$

© 2022 Sung-Pil Hong. All rights reserved.

## Proposition 2.3

*Suppose* $\bar{x}$ *is a local optimum of* $\max\{f(x) \,|\, g(x) \leq 0\}$. *Let* $A(\bar{x})$ *be the indices of active constraints of* $\bar{x}$. *If a regularity condition is satisfied, namely that* $\{\nabla g_i(\bar{x}) : i \in A(\bar{x})\}$ *are linearly independent, then there is* $\lambda \in \mathbb{R}^m$ *such that*

$$\begin{aligned} 0 &= \nabla f(\bar{x}) - \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}), \\ \lambda &\geq 0, \\ g(\bar{x})^T \lambda &= 0 \quad \textit{or equivalently, } \lambda_i = 0, \forall i \notin A(\bar{x}). \end{aligned} \tag{2.8}$$

## Exercise 2.4

*Restate the necessary condition for* $\min\{f(x) : g_i(x) \geq 0,\ 1 \leq i \leq m\}$.

## Exercise 2.5

*Repeat for* $\min\{f(x) : g_i(x) \leq 0,\ 1 \leq i \leq m\}$ *and* $\max\{f(x) : g_i(x) \geq 0,\ 1 \leq i \leq m\}$.

© 2022 Sung-Pil Hong. All rights reserved.

## Example 2.6

*Find every local minimum and maximum of* $f(x) = x_1^2 - x_2^2$ *over the set of* $x$ *satisfying* $g(x) = x_1^2 + x_2^2 - 1 \leq 0$.

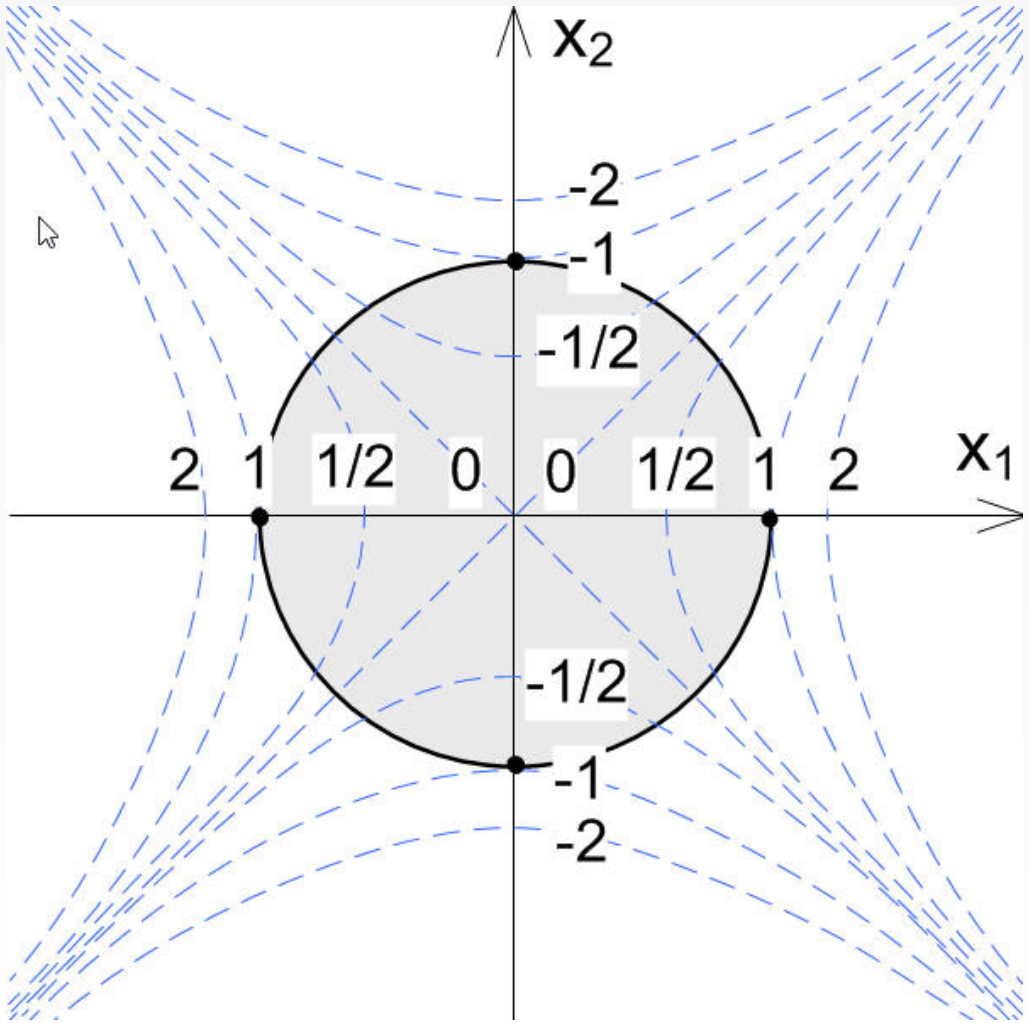
Since $(0, 0)$ is not a local optimum, every minimum or maximum has a nonzero gradient $\nabla g(x)$. Hence the regularity condition holds and they are necessarily a KKT point: there is $\lambda \geq 0$ $(\leq 0)$ such that $\nabla f(x) - \lambda \nabla g(x) = 0$ for a local maximum (minimum, resp. ).

$$\nabla f(x) - \lambda \nabla g(x) = \begin{bmatrix} 2x_1 \\ -2x_2 \end{bmatrix} - \lambda \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 0. \tag{2.9}$$

From (2.9) and $x_1^2 + x_2^2 = 1$, we get two groups of KKT points: $(+1, 0)$ and $(-1, 0)$ for $\lambda = 1$ and $(0, +1)$ and $(0, -1)$ for $\lambda = -1$. In fact, the first two points are maximizers and the last two points are minimizers.

© 2022 Sung-Pil Hong. All rights reserved.

To see this, we need to rely on extra information. In the following sketch of the optimization problem, we can confirm that they are indeed maxima and minima.




© 2022 Sung-Pil Hong. All rights reserved.

Example 2.7

*Consider the following optimization problem:*

$$\begin{aligned} \max \quad & \log x_1 + \log x_2 \\ \textit{sub. to} \quad & x_1 + 2x_2 \leq 20 \\ & x_1 \leq 8 \end{aligned}$$

Let's solve this problem analytically. The gradients $[1, 2]$ and $[1, 0]$ of two constraints are linearly independent. By KKT necessity, the optimal value of $x$ must satisfy the KKT conditions. Therefore, if we find *all* the values of $x$ and $\lambda$, the optimum must exist among them. Let $\lambda_1$ and $\lambda_2$ denote the Lagrange multipliers for the first and second constraint, respectively.

© 2022 Sung-Pil Hong. All rights reserved.

There are four cases:

1. $\lambda_1 = \lambda_2 = 0$. Then we must have $\nabla f(x) = [1/x_1, 1/x_2]^T = 0$. There is **no feasible solution,** because ....
2. $\lambda_1 > 0, \lambda_2 = 0$. Then we must have $\nabla f(x) = [1/x_1, 1/x_2]^T = \lambda_1 \nabla g_1(x) = \lambda_1 [1, 2]^T$ and $g_1(x) = x_1 + 2x_2 = 20$. Solving these equations, we find $x_1 = 10, x_2 = 5$, and $\lambda_1 = 1/10$, which is **infeasible** because $x_1 > 8$.
3. $\lambda_1 = 0, \lambda_2 > 0$. Then we must have $\nabla f(x) = [1/x_1, 1/x_2]^T = \lambda_2 \nabla g_2(x) = \lambda_2 [1, 0]^T$ and $g_2(x) = x_1 = 8$. There is **no feasible solution,** because ....
4. $\lambda_1 > 0, \lambda_2 > 0$. Then we must have $\nabla f(x) = [1/x_1, 1/x_2]^T = \lambda_1 \nabla g_1(x) + \lambda_2 \nabla g_2(x) = \lambda_1 [1, 2]^T + \lambda_2 [1, 0]^T$, and $g_1(x) = x_1 + 2x_2 = 20$ and $g_2(x) = x_1 = 8$. Solving, we find $x_1 = 8, x_2 = 6, \lambda_1 = 1/12$, and $\lambda_2 = 1/24$, which is **feasible.**

We conclude that $x^* = (8, 6)$ **is the optimal solution**, since it is the only point that satisfies the KKT conditions.

© 2022 Sung-Pil Hong. All rights reserved.

## Exercise 2.8

**Smallest enclosing sphere** *Consider a set of points $p_1 \cdots p_m$ in $\mathbb{R}^n$. What is the smallest sphere centered at the origin that encloses all of these points?*

*Let $x$ denote the radius of the sphere. Then $p_j$ is in the sphere if and only if $\|p_j\|_2 \leq x$, and the optimization problem is*

$$\begin{aligned} \min \quad & x \\ \text{sub. to} \quad & x \geq \|p_j\|_2, \quad j = 1 \ldots m \end{aligned}$$

*Each of the $j$ constraints is linear. Under the assumption that $\|p_j\|_2 \neq \|p_k\|_2$, let $j^* = \arg\max_j \{\|p_j\|_2\}$. The optimal solution is given by $x^* = \|p_{j^*}\|_2$ (why?).*

1. *At $x^*$, there is only one binding constraint. What is it?*
2. *Determine the values of the Lagrange multipliers $\lambda_1 \cdots \lambda_m$ at $x^*$ and show that they satisfy the KKT conditions.*

© 2022 Sung-Pil Hong. All rights reserved.

## Exercise 2.9

$$\begin{aligned} \min \quad & 3x_1^2 + 3x_2^2 - 2x_1x_2 - 2x_1 + 6x_2 + 3 \\ \textit{sub. to} \quad & -x_1 + x_2 \geq -1 \\ & (x_1 - 2)^2 + x_2^2 \leq 1 \end{aligned} \tag{2.10}$$

© 2022 Sung-Pil Hong. All rights reserved.

Now we explore optimality condition for the equality constrained cases.

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) = \begin{bmatrix} g_1(x) \\ \cdots \\ g_m(x) \end{bmatrix} = 0 \end{aligned} \tag{2.11}$$

**Proposition 2.10**

*Suppose $x^*$ is a local optimum of (2.11). Assume $\nabla g_i(x^*)$'s are linearly independent. Then there is $\lambda^*$: $\nabla f(x^*) = \lambda_1^* \nabla g_1(x^*) + \cdots + \lambda_m^* \nabla g_m(x^*)$.*

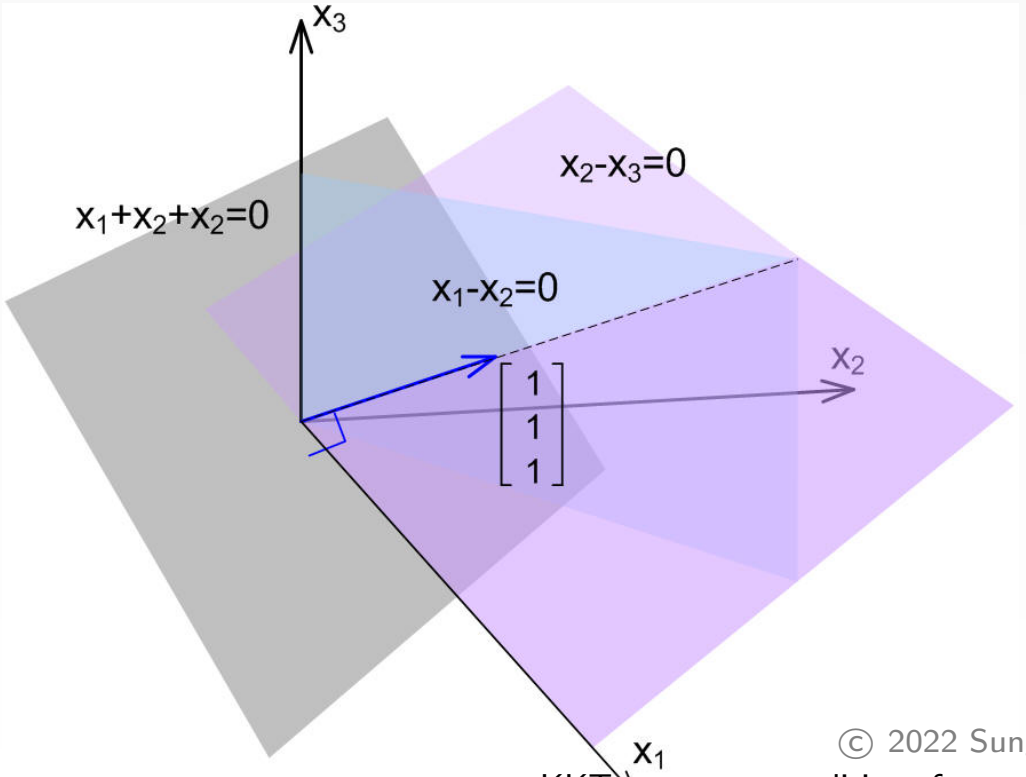
Although we are doing it informally, it will introduce us to an alternative proof of KKT conditions and deepen our insight on the optimality structure. To do so, we first make some observations on the row and null spaces of a matrix.

© 2022 Sung-Pil Hong. All rights reserved.

Consider the set of the solutions of (2.12), called the *null space* of $A$.

$$Ax = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.12}$$

The null space is exactly the set of vectors orthogonal to the row vectors of $A$, hence also to their linear combinations, called the *row space* of $A$. By solving (2.12), it is the set of multiples of vector $[1, 1, 1]^T$, the intersection of the blue and red planes in the picture.




© 2022 Sung-Pil Hong. All rights reserved.

Hence the dimension of the null space is $1$. The dimension of the row space is $2$ since $[1, -1, 0]^T$ and $[0, 1, -1]^T$ are independent. The two spaces are orthogonal to each other and the sum of their dimensions is equal to $3$, the dimension of the whole space. We call this rank-and-nullity theorem.

Since the row space are orthogonal to vector $[1, 1, 1]$, the it is included in $\{x : x_1 + x_2 + x_3 = 0\}$, the null space of the matrix $B = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$, the grey plane. Since the row space has two independent vectors, it is the same as the two dimensional grey plane, the null space of $B$.

We have observed the following.

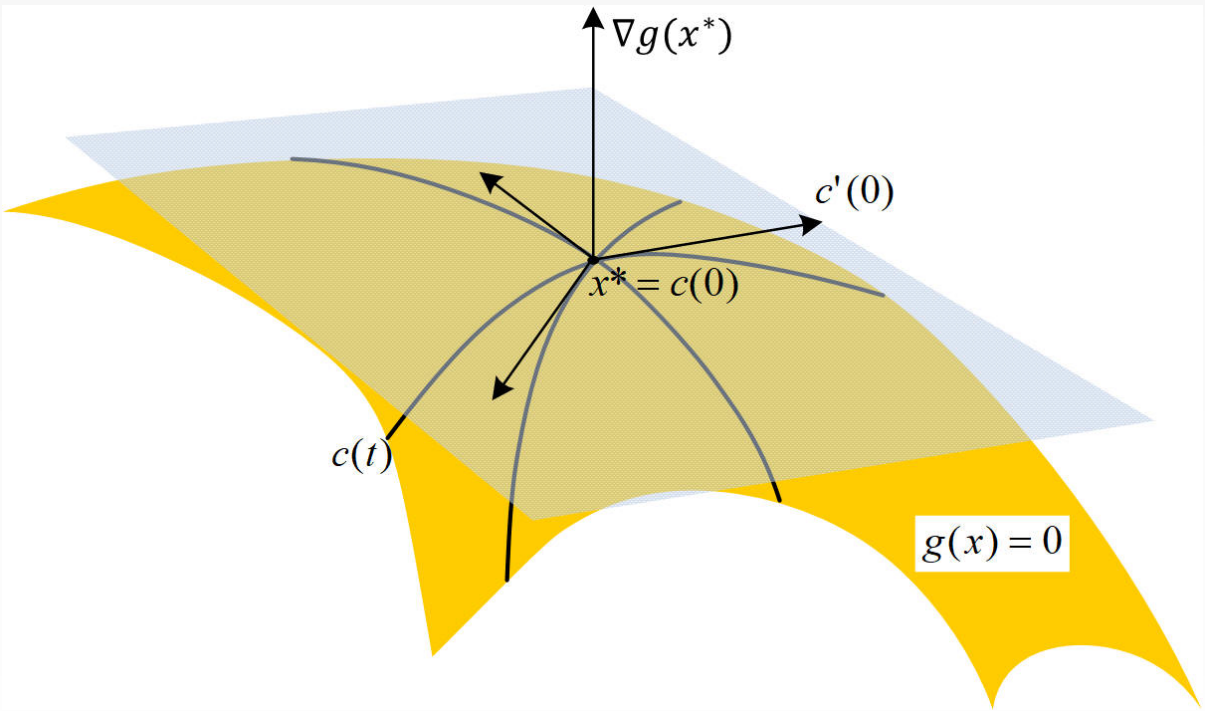
**Lemma 2.11**

*The set of vectors orthogonal to the null space of a matrix $A$ is the row space of $A$. Namely, the null space of the null space of $A$ is the row space of $A$.*

Bookmark this lemma in your mind as it will be used in the (sketch of) proof of Proposition 2.10.

© 2022 Sung-Pil Hong. All rights reserved.

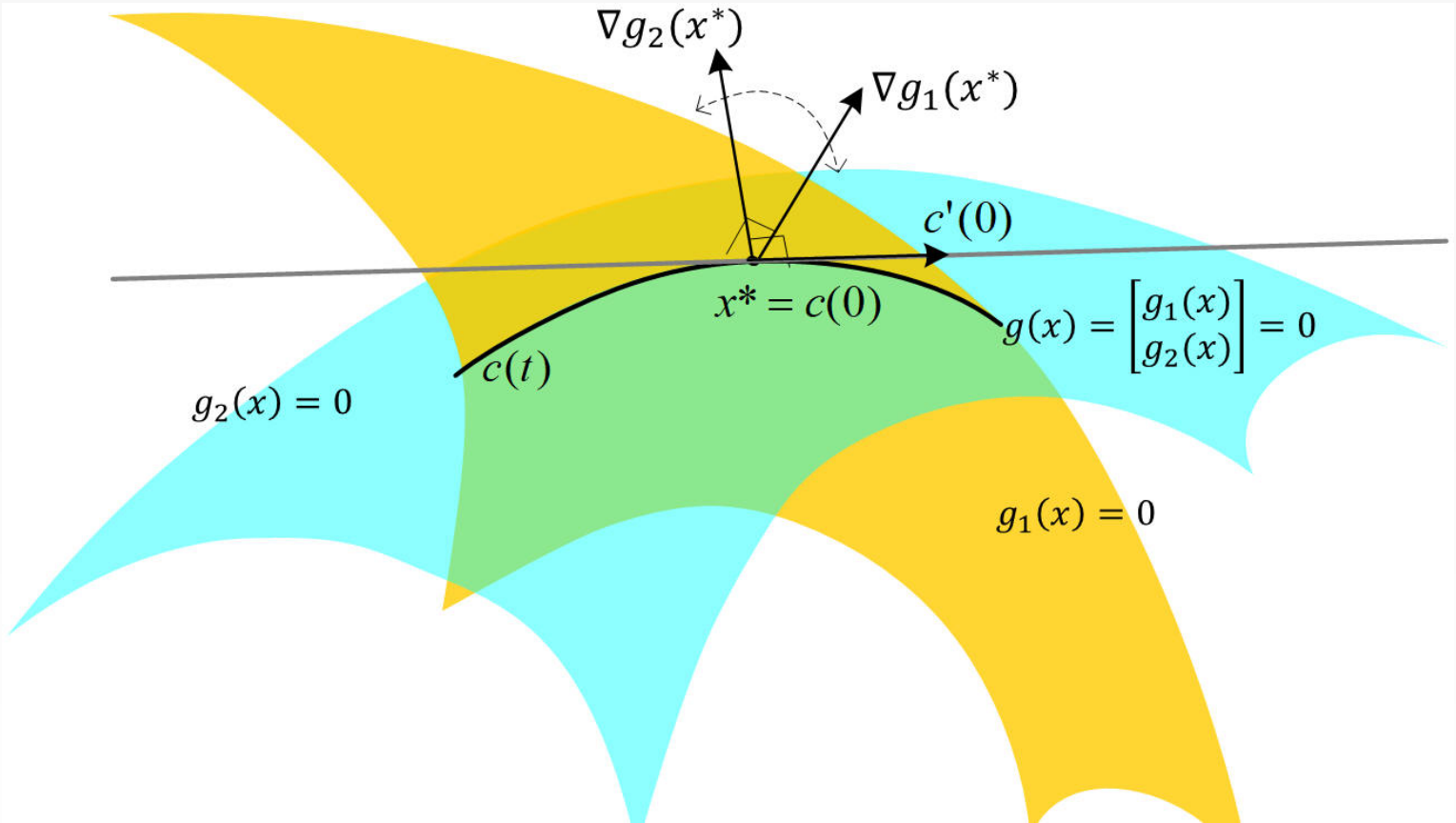
Let $c : \mathbb{R} \to \mathbb{R}^n$ be a differentiable curve imbedded in the solution set $g(x) = 0$, i.e. $g(c(t)) = 0 \; \forall \; t$ with $c(0) = x^*$. Let $m = 1$ as in the picture. Then $0 = (g(c(0)))' = \nabla g(c(0))^T c'(0) = \nabla g(x^*)^T c'(0)$. Notice that there are an infinite number of ways of arranging an imbedded curve on the surface $g(x) = 0$ passing through $x^*$. It suggests $c'(0)$ of such curves covers the supporting hyperplane perpendicular to $\nabla g(x^*)$.




We call the set of $y = c'(0)$ of such curves the *tangent plane* to the surface at $x^*$ which is in fact the null space $\{y : \nabla g(x^*)^T y = 0\}$.

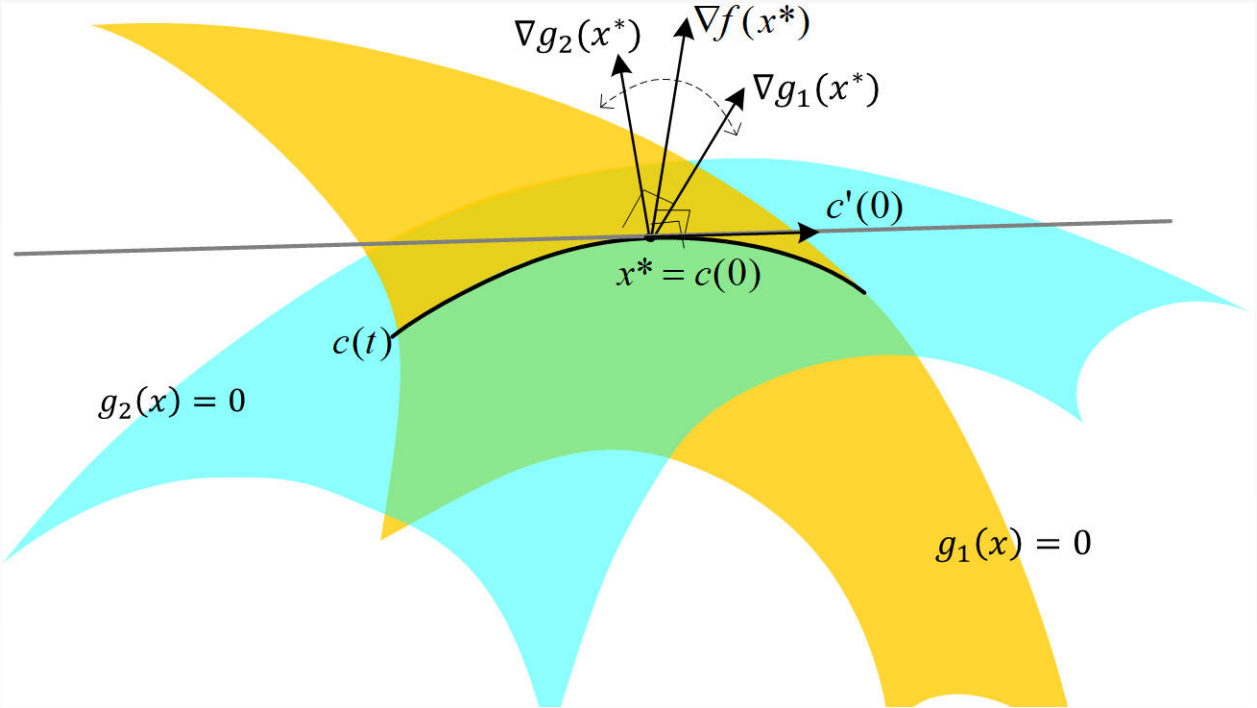
© 2022 Sung-Pil Hong. All rights reserved.

In general, the tangent plane to the set $\left\{ x : g(x) = \begin{bmatrix} g_1(x) \\ \dots \\ g_m(x) \end{bmatrix} = 0 \right\}$ is the null space of $Dg(x^*)$, the $y$'s such that $\nabla g_i(x^*)^T y = 0$, $1 \leq i \leq m$, the grey line in the next picture illustrating a case when $m = 2$.




© 2022 Sung-Pil Hong. All rights reserved.

Since $x^* = c(0)$ is a local minimum, $t = 0$ is a local minimum of $h(t) := f(c(t))$. Hence $h'(0) = \nabla f(c(0))^T c'(0) = \nabla f(x^*)^T c'(0) = 0$. Thus $\nabla f(x^*)$ is perpendicular to every vector of the tangent plane, i.e. belongs to the null space of the tangent plane which is the null space of the matrix $Dg(x^*)$.



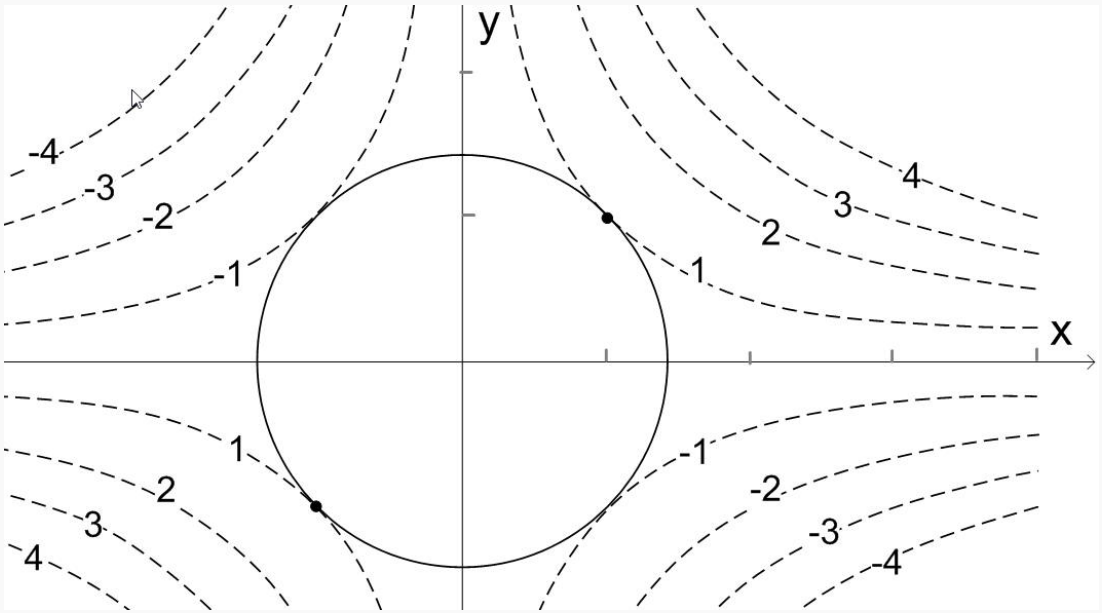

Hence from Lemma 2.11, $\nabla f(x^*)$ belongs to the row space of $Dg(x^*)$ and therefore a linear combination of $\nabla g_i(x^*)$'s.

$$\exists \lambda \in \mathbb{R}^m : \nabla f(x^*) = \lambda_1 \nabla g_1(x^*) + \cdots + \lambda_m \nabla g_m(x^*). \quad \square$$

© 2022 Sung-Pil Hong. All rights reserved.

## Example 2.12

*Consider the problem* $\max\{xy : x^2 + y^2 = 1\}$ *and the KKT condition* $\begin{bmatrix} y \\ x \end{bmatrix} - \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix} = 0$ *which implies* $y = 2\lambda x$ *and* $x = 2\lambda y$. *From the constraint, we get* $\lambda = \pm\frac{1}{2}$, *and the KKT points* $(1, 1)$, $(-1, -1)$, $(1, -1)$, *and* $(-1, 1)$. *As in the figure below the first two are maxima.*




© 2022 Sung-Pil Hong. All rights reserved.

## Exercise 2.13

*Solve the following problem.*

1.
$$\begin{aligned} \min \quad & 2x_1^2 + 2x_1x_2 + x_2^2 \\ \textit{sub. to} \quad & 2x_1 + x_2 = 4. \end{aligned}$$

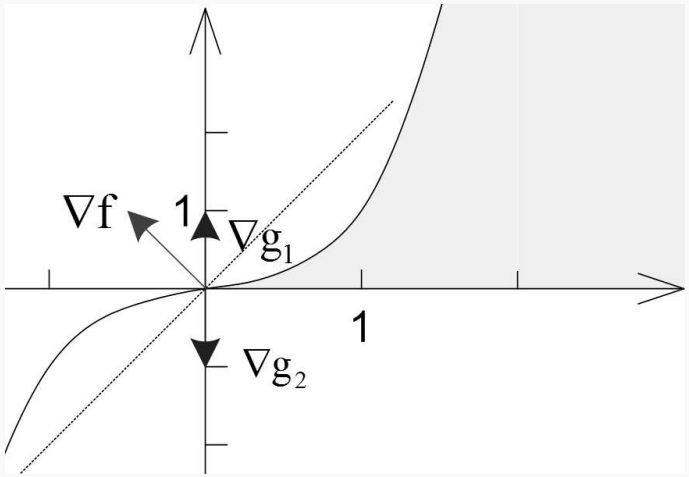
2.
$$\begin{aligned} \min \quad & 3x_1^2 + 2x_1x_2 + x_2^2 - 2x_1 - 2x_2 + 3 \\ \textit{sub. to} \quad & -2x_1 + 5x_2 = 12 \end{aligned}$$

© 2022 Sung-Pil Hong. All rights reserved.

## Remark 2.14

- The KKT conditions do not guarantee local optimality: consider the feasible solution $(0, 0)$ of $\min\{x_2 : -x_1^3 + x_2 \geq 0\}$.
- The following example shows that the 'regularity condition' is necessary. (Without it, there may be no $y : \nabla g_i(\bar{x})^T y < 0$.)

$$\begin{aligned} \max \quad & -x_1 + 2x_2 \\ \text{s.t.} \quad & -x_1^3 + x_2 \leq 0 \\ & -x_2 \leq 0. \end{aligned}$$

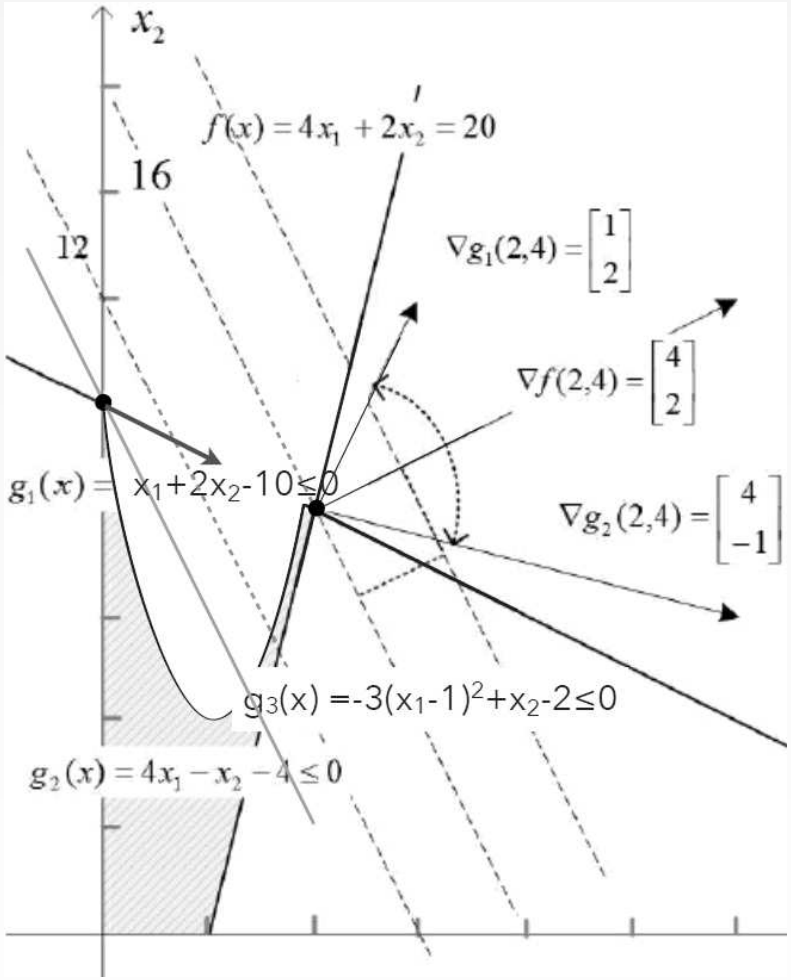



- We can extend the KKT conditions for the optimization problems with both inequality and equality constraints.

© 2022 Sung-Pil Hong. All rights reserved.

We need an additional condition for a local optimum to be (global) optimal.

$$
\begin{array}{rrll}
\max & 4x_1 & +2x_2 & \\
\text{sub. to} & x_1 & +2x_2 - 10 & \leq 0 \\
& x_1^2 & -x_2 & \leq 0 \\
& -3(x_1-1)^2 & +x_2 - 2 & \leq 0 \\
& & x \geq 0 &
\end{array}
$$




The feasible solution $(0, 5)$ is a local optimum but not a (global) optimum.

© 2022 Sung-Pil Hong. All rights reserved.