[SWCON253] Machine Learning – Lec.14a

Constrained Optimization

김 휘 용

hykim.v@khu.ac.kr

경희대학교
KYUNG HEE UNIVERSITY

Contents

References

- *Chapter 7, Mathematics for Machine Learning* by Deisenroth, Faisal, and Ong (https://mml-book.com)
- *Intro to Deep Learning & Generative Models* by Sebastian Raschka (http://pages.stat.wisc.edu/~sraschka/teaching/stat453-ss2020/)
- *패턴 인식* by 오일석, *기계 학습* by 오일석

NOTE:

All the derivations in this lecture assume ***convexity*** of the variables & functions.

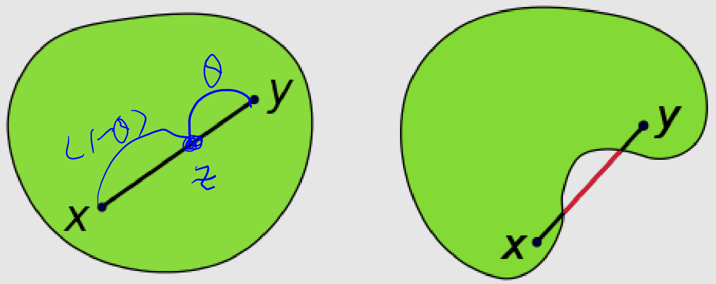
Convex Set & Convex Function

Convex Set

Definition 7.2. A set $\mathcal{C}$ is a convex set if for any $x, y \in \mathcal{C}$ and for any scalar θ with $0 \leqslant \theta \leqslant 1$, we have

$$z = \theta x + (1-\theta)y \in \mathcal{C}\,. \tag{7.29}$$

Linear segment btw two points in C lies in C.



Convex Function

Definition 7.3. Let function $f : \mathbb{R}^D \to \mathbb{R}$ be a function whose domain is a convex set. The function f is a convex function if for all $\boldsymbol{x}, \boldsymbol{y}$ in the domain of f, and for any scalar θ with $0 \leqslant \theta \leqslant 1$, we have

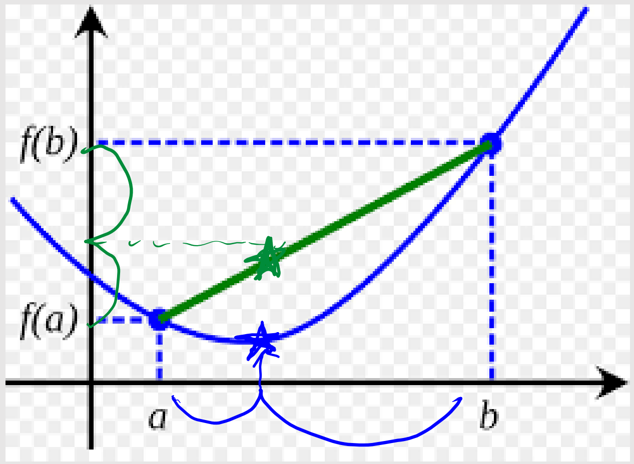
Jensen's Inequality

$$\boxed{f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leqslant \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y})}\,. \tag{7.30}$$

Remark. A concave function is the negative of a convex function. ◇

☆ ≤ ☆

A straight line between any two points of the function lie above the function.



(a ⇒ x, b ⇒ y)

☆ : 내분점의 함숫값
☆ : 함숫값의 내분

Convex Set & Convex Function (cont'd)

Properties of Convexity

1. If $\mathcal{S}$ is convex and $\alpha \in \Re$, then the set $\alpha\mathcal{S}$ is convex.
2. If $\mathcal{S}_1, \mathcal{S}_2$ are convex sets, then the set $\mathcal{S}_1 + \mathcal{S}_2 = \{\mathbf{y} \mid \mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}$ is convex.
3. The intersection of convex sets is convex.
4. The union of convex sets is usually nonconvex.
5. If f is a convex function and $\alpha \geq 0$, then αf is a convex function.
6. If f_1, f_2 are convex on a set $\mathcal{S}$, then $f_1 + f_2$ is also convex on $\mathcal{S}$.

Testing Convexity

A differentiable function is convex,
if and only if its **Hessian is positive-semidefinite** in its entire convex domain.

1. *Constrained* Optimization Problems

With *No* Constraints

$$\boxed{\text{minimize } f(x)}$$

- *Unconstrained* Optimization
- Solution:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \mathbf{0}$$

With *Equality* Constraints

$$\boxed{\begin{array}{l}\text{minimize } f(\underset{\sim}{x}) \\ \text{subject to } g(\underset{\sim}{x}) = 0\end{array}}$$

- Example

$$f(x_1, x_2) = x_1^2 + x_2^2 - 1$$
$$g(x_1, x_2) = x_1 + x_2 - 1$$

With *Inequality* Constraints

$$\boxed{\begin{array}{l}\text{minimize } f(\underset{\sim}{x}) \\ \text{subject to } g(\underset{\sim}{x}) \le 0\end{array}}$$

- Example

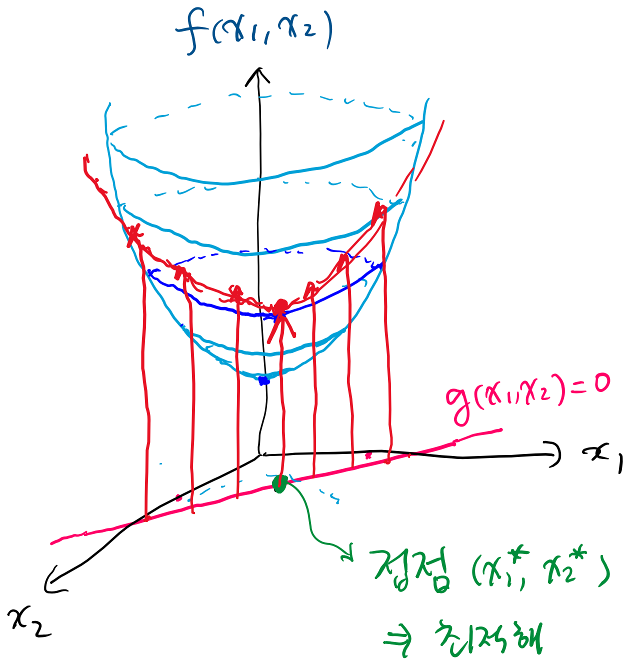
$$\text{maximize}\, x \cdot y$$
$$\text{subject to } 2x + 2y \le 1$$

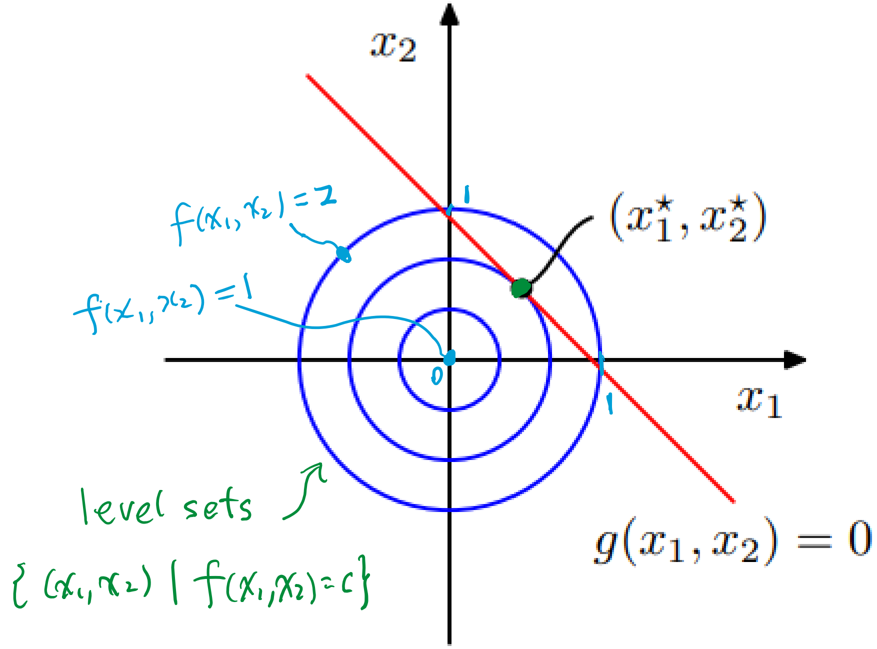
$$\text{minimize}(-x \cdot y)$$
$$\text{subject to } 2x + 2y - 1 \le 0$$

◆ An Illustrative Example

minimize $f(\underset{\sim}{x})$
subject to $g(\underset{\sim}{x}) = 0$

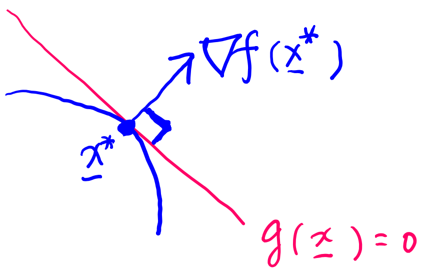
$$f(x_1, x_2) = x_1^2 + x_2^2 + 1$$
$$g(x_1, x_2) = x_1 + x_2 - 1$$



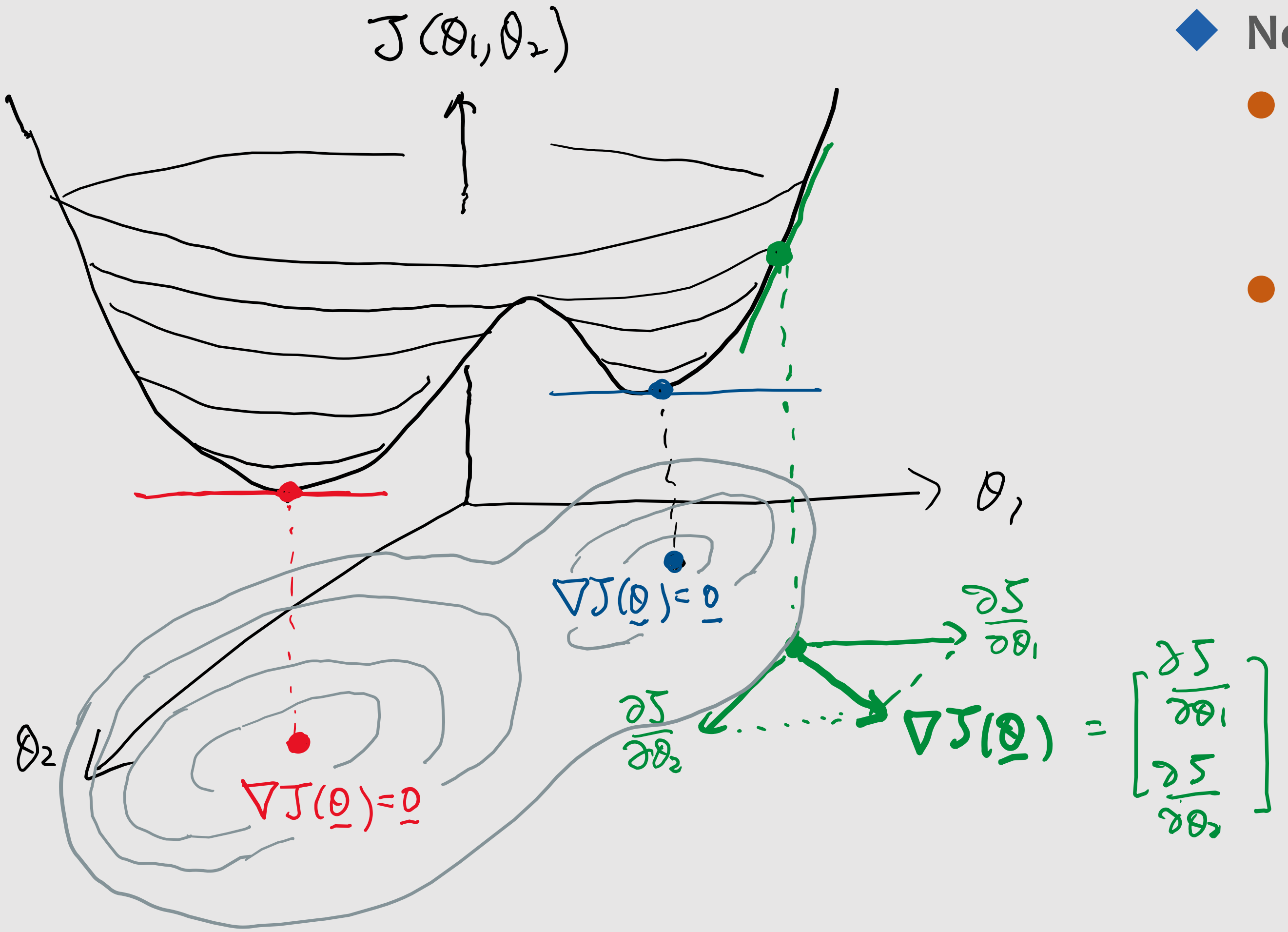


At the optimal point $\boldsymbol{x}^*$, the gradient of f is orthogonal to the surface $\{g(\boldsymbol{x}) = 0\}$.

Consequently, $\nabla g(\boldsymbol{x})$ & $\nabla f(\boldsymbol{x})$ are parallel at $\boldsymbol{x}^*$.



(Revisit) Gradient of a Function



- **Note**
  - Gradient vector indicates the ***direction*** of steepest ascent of the loss function *J*. ***on the input (parameter) space***.
  - Thus, for the left example, the gradient of *J* is a 2D vector *on the 2D theta-plane*. I.e., it is *not on the 3D loss surfaces*. (Refer to the *green drawings* on the left.)

2. Lagrange Multiplier for *Equality* Constraints (cont'd)

Equality Constrained Optimization

$$\begin{aligned}&\text{minimize } f(\underset{\sim}{x})\\ &\text{subject to } g(\underset{\sim}{x}) = 0\end{aligned}$$

- At the optimal point $\boldsymbol{x}^*$, $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ are <u>*parallel*</u>.
  - ★ Hence, there exists some $\lambda \in \boldsymbol{R}$ such that $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \mathbf{0}$.
- We define the ***Lagrangian*** function $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$,

  where $\lambda \in \boldsymbol{R}$ is called ***Lagrangian multiplier***.
- Now observe that:
  - ★ $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \mathbf{0} \;\Leftrightarrow\; \nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}$
  - ★ $g(\boldsymbol{x}) = 0 \;\Leftrightarrow\; \nabla_{\lambda} L(\boldsymbol{x}, \lambda) = 0$

<u>Example</u>

$$f(x_1, x_2) = x_1^2 + x_2^2 - 1$$
$$g(x_1, x_2) = x_1 + x_2 - 1$$

$$\begin{aligned}L(\boldsymbol{x}, \lambda) &= x_1^2 + x_2^2 - 1\\ &\quad + \lambda(x_1 + x_2 - 1)\end{aligned}$$

$$\nabla_{x_1} L(\boldsymbol{x}, \lambda) = 2x_1 + \lambda = 0$$
$$\nabla_{x_2} L(\boldsymbol{x}, \lambda) = 2x_2 + \lambda = 0$$
$$\nabla_{\lambda} L(\boldsymbol{x}, \lambda) = x_1 + x_2 - 1 = 0$$

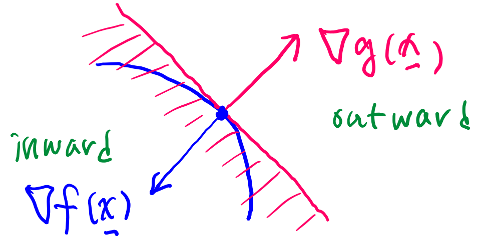
$$\therefore (x_1^*, x_2^*) = (0.5, 0.5)$$

3. Lagrange Multiplier for *Inequality* Constraints

Inequality Constrained Optimization

minimize $f(\underline{x})$
subject to $g(\underline{x}) \le 0$

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$



- **<u>Case 1</u>**: Constraint is "***active***" (i.e., $\boldsymbol{x}^*$ is on $g(\boldsymbol{x}) = 0$)
  - ★ When $\boldsymbol{x}_u^*$ is outside the boundary: $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ points <u>outwards</u> and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ points <u>inwards</u>.
    → $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \mathbf{0}$ for some $\lambda > 0$.

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}$$

- **<u>Case 2</u>**: Constraint is "***inactive***" (i.e., $\boldsymbol{x}^*$ is **not** on $g(\boldsymbol{x}) = 0$)
  - ★ In this case, we can treat the problem as ***unconstrained*** optimization ($\lambda = 0$).
  - ★ Thus, we have $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \mathbf{0}$ at the optimal point $\boldsymbol{x}^*$.

$\Rightarrow$
- Active $(g(\underline{x}) = 0)$: $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}$ for some $\lambda > 0$.
- Inactive $(g(\underline{x}) < 0)$: $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}$ for $\lambda = 0$.

$\Rightarrow$

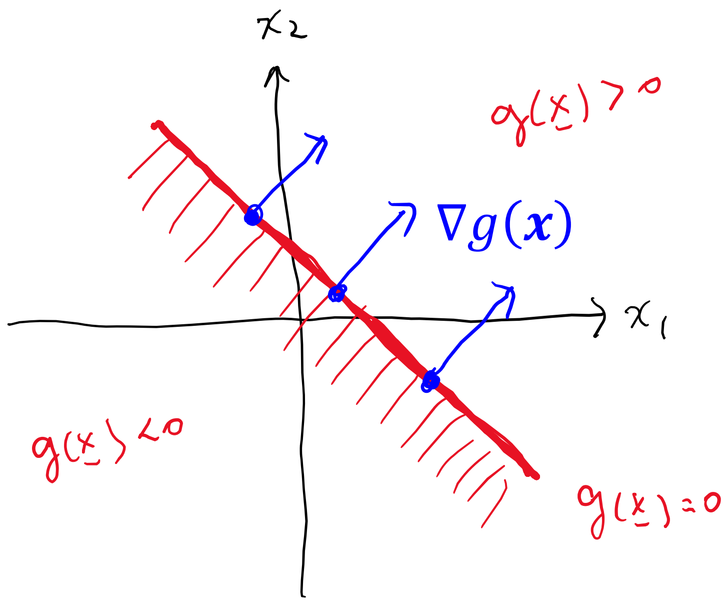
$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}$$

- $g(\boldsymbol{x}) = 0,\ \lambda > 0$ (*active*)
- $g(\boldsymbol{x}) < 0,\ \lambda = 0$ (*inactive*)

3. Lagrange Multiplier for *Inequality* Constraints (cont'd)

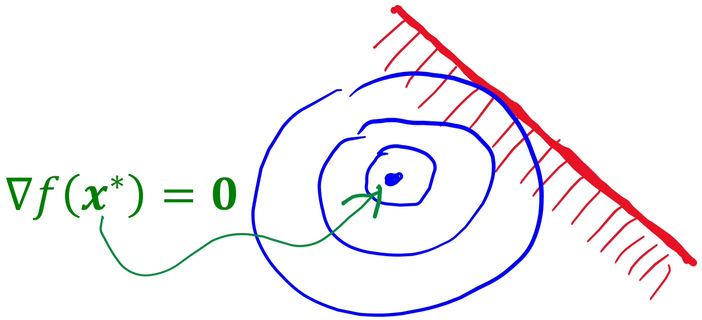
[Note 1] Gradient of g

- Inequality constraint: $g(\boldsymbol{x}) \leq 0$
- $\nabla g(\boldsymbol{x})$ is *outward* the boundary
  - ★ since at the boundary $g(\boldsymbol{x})$ increases on the outward direction (from negative to positive).

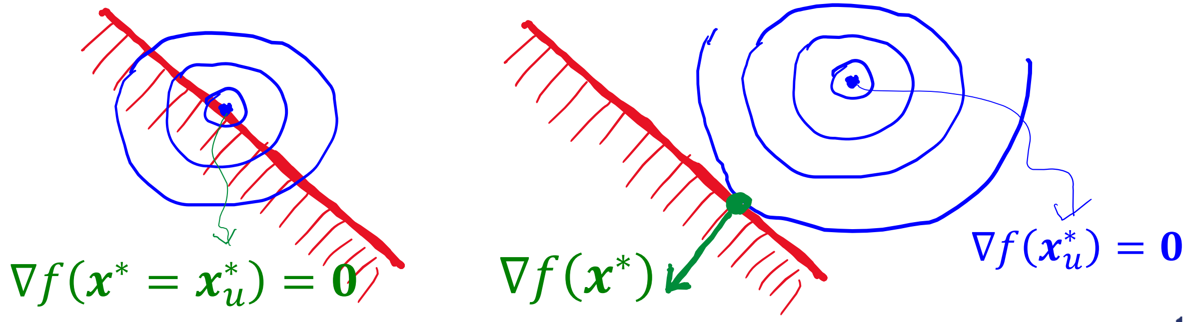


[Note 2] Gradient of f

- **Inactive case** (i.e., solution is *inside* the boundary)
  - ★ The unconstrained solution must reside *inside* the boundary → $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ $(\boldsymbol{x}^* = \boldsymbol{x}^*_{unc})$



- **Active case** (i.e., solution is *on* the boundary)
  - ★ The unconstrained solution must reside either *on* or *outside* the boundary
    → $\nabla f(\boldsymbol{x}^*)$ is either ***zero*** or ***inward*** the boundary.



3. Lagrange Multiplier for *Inequality* Constraints (cont'd)

Inequality Constrained Optimization (cont'd)

- We can summarize both cases using the Lagrangian again.

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

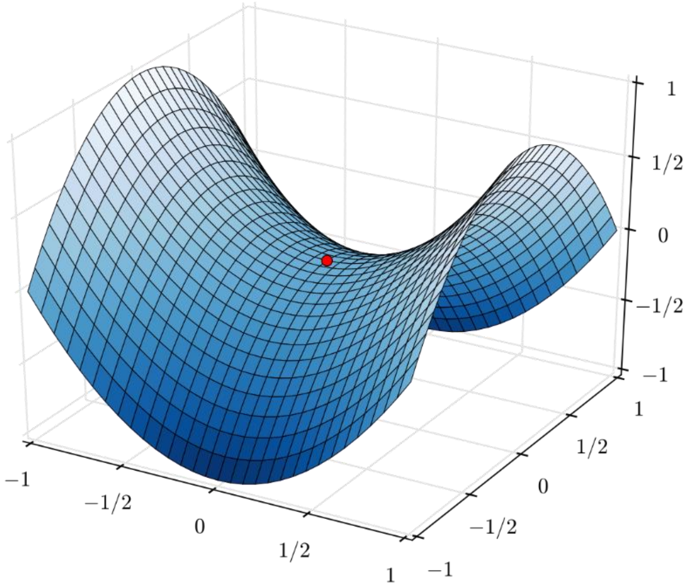
$\boxed{\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = \mathbf{0}}$	(stationarity or saddle point condition)
$g(\boldsymbol{x}) \leq 0$	(primal feasibility)
$\lambda \geq 0$	(dual feasibility)
$\lambda\, g(\boldsymbol{x}) = 0$	(complementary slackness)

check

- $\boxed{g(\boldsymbol{x}) = 0,}\ \lambda > 0$ *(active)* $\equiv \boxed{\nabla_{\lambda} L(\boldsymbol{x}, \lambda) = 0}$
- $g(\boldsymbol{x}) \leq 0,\ \boxed{\lambda = 0}$ *(inactive)*

check

A ***saddle point*** on the graph of $z=x^2-y^2$ (hyperbolic paraboloid)



https://en.wikipedia.org/wiki/Saddle_point#/media/File:Saddle_point.svg

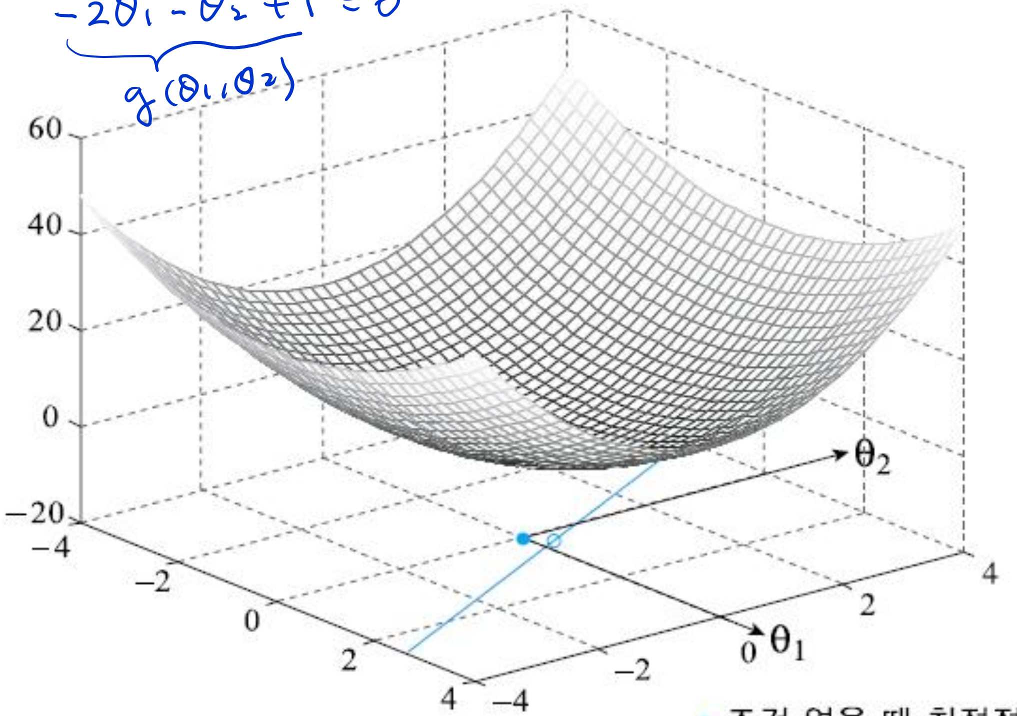
- The conditions in the yellow box is called the ***Karush-Kuhn-Tucker (KKT) condition***

★ It is sometimes also called as *first-order necessary condition*.

<> Examples – *Equality Constraint*

- 등식 조건부 최적화 문제: $J(\boldsymbol{\theta}) = \theta_1^{\ 2} + 2\theta_2^{\ 2}$
  - 조건이 없을 때 최소 점은 $(0,0)^T$
  - '$2\theta_1 + \theta_2 = 1$' 이라는 조건 하에 최소점을 구하면?

$\underbrace{-2\theta_1 - \theta_2 + 1}_{g(\theta_1, \theta_2)} = 0$



- 조건 없을 때 최적점$(0, 0)^T$
- 조건 있을 때 최적점$(4/9, 1/9)^T$

그림 11.11 등식 조건부 최적화 문제의 예

라그랑제 함수: $L(\boldsymbol{\theta}, \lambda_1) = (\theta_1^{\ 2} + 2\theta_2^{\ 2}) - \lambda_1 \overbrace{(2\theta_1 + \theta_2 - 1)}^{g(\theta_1, \theta_2)}$

$\boldsymbol{\theta}$로 미분한 식을 $\mathbf{0}$으로 둠:

$$\begin{cases} \dfrac{\partial L}{\partial \theta_1} = 2\theta_1 - 2\lambda_1 = 0 \\ \dfrac{\partial L}{\partial \theta_2} = 4\theta_2 - \lambda_1 = 0 \end{cases}$$

- $\boldsymbol{\lambda}$로 미분한 식을 $\mathbf{0}$으로 둠: $\dfrac{\partial L}{\partial \lambda_1} = -(2\theta_1 + \theta_2 - 1) = 0$

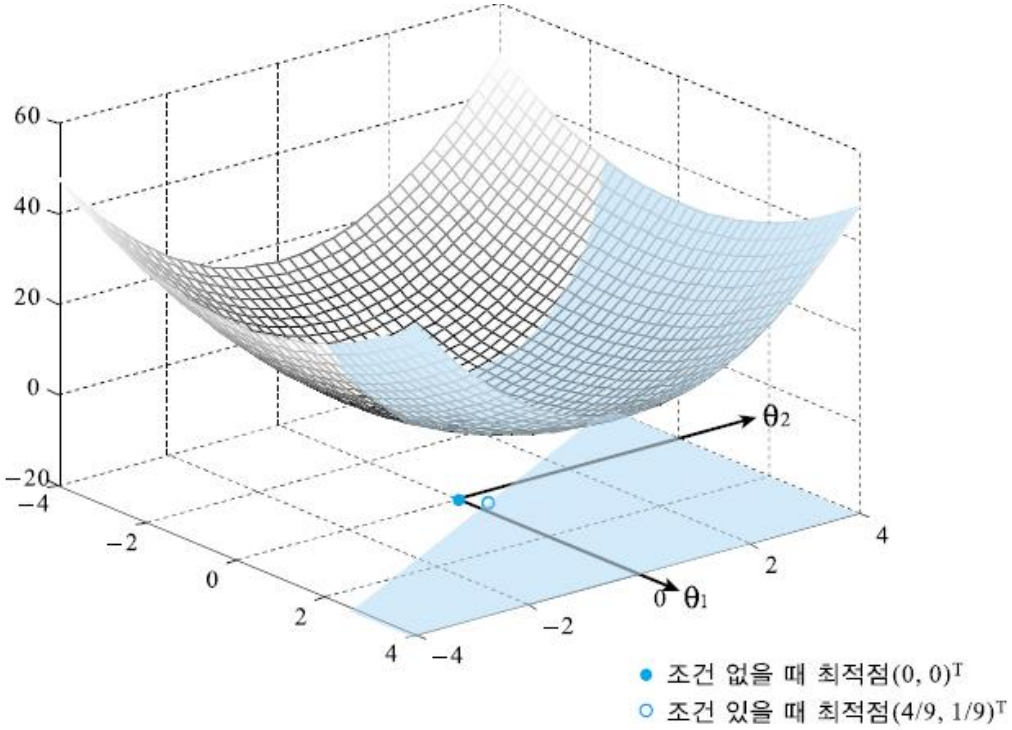
이 식을 풀면, $\hat{\boldsymbol{\theta}} = (\frac{4}{9}, \frac{1}{9})^T$, $\lambda = \frac{4}{9} > 0$

<> Examples – *Inequality Constraint*

◆ 부등식 조건부 최적화 문제:

$$J(\boldsymbol{\theta}) = \theta_1^{\,2} + 2\theta_2^{\,2}$$

● '$2\theta_1 + \theta_2 \geq 1$'이라는 조건 하에 최소점을 구하라.



(a) $f_1(\boldsymbol{\theta}) = 2\theta_1 + \theta_2 - 1 \geq 0$

$g(\theta)$

라그랑제 함수: $L(\boldsymbol{\theta}, \lambda_1) = (\theta_1^{\,2} + 2\theta_2^{\,2}) - \lambda_1 (2\theta_1 + \theta_2 - 1)$

KKT 조건:
$$\begin{cases} \dfrac{\partial L}{\partial \theta_1} = 2\theta_1 - 2\lambda_1 = 0 \\ \dfrac{\partial L}{\partial \theta_2} = 4\theta_2 - \lambda_1 = 0 \\ \lambda_1 \geq 0 \\ \lambda_1 (2\theta_1 + \theta_2 - 1) = 0 \end{cases}$$

이제 KKT 조건을 풀어 이를 만족하는 해 $\hat{\boldsymbol{\theta}}$를 구하면 된다. 이 예는 라그랑제 승수를 하나만 가지므로 쉽게 풀 수 있다. 마지막 식에서 $\lambda_1 = 0$이거나 $2\theta_1 + \theta_2 - 1 = 0$이어야 한다. 먼저 $\lambda_1 = 0$이라고 가정해 보자. 그럼 $\theta_1 = \theta_2 = 0$이 되어 주어진 조건 $f_1(\boldsymbol{\theta}) = 2\theta_1 + \theta_2 - 1 \geq 0$을 만족하지 못한다. 두 번째 경우를 가지고 풀어 보면 $\hat{\boldsymbol{\theta}} = (\frac{4}{9}, \frac{1}{9})^{\mathrm{T}}$을 얻는다.

① active : $\boxed{\nabla_\lambda L = 0} \Rightarrow -2\theta_1 - \theta_2 + 1 = 0$

Solution : $\theta = (\frac{4}{9}, \frac{1}{9})^T$, $\lambda = \frac{4}{9} > 0$. (O)

② inactive : $\boxed{\lambda = 0} \Rightarrow \theta = 0 \Rightarrow g(\theta) = 1 > 0$ (X)

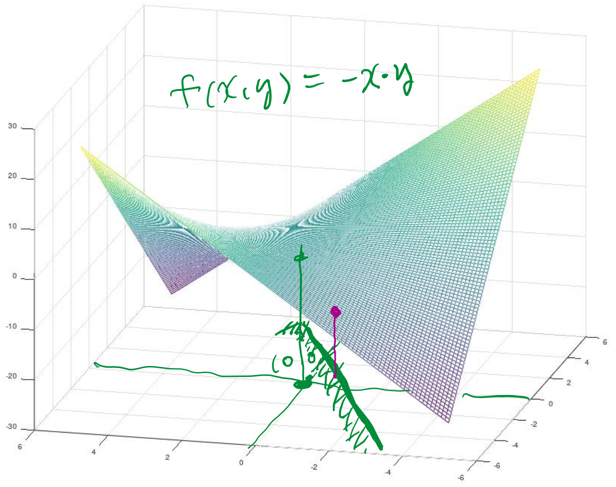
Weird Example? – *Non-Convex Cost*

- **Non-Convex Cost의 경우**

non-convex

$$\text{minimize}(-x \cdot y)$$
subject to $2x + 2y - 1 \le 0$

```
>> tx = ty = linspace(-5, 5, 100)';
>> [x, y] = meshgrid(tx, ty);
>> f = -x.*y;
>> mesh(tx, ty, f)
```



$$L(x, y, \lambda) = -xy + \lambda(2x + 2y - 1)$$

$$\begin{cases} \partial L/\partial x = -y + 2\lambda \overset{!}{=} 0 & ① \\ \partial L/\partial y = -x + 2\lambda \overset{!}{=} 0 & ② \end{cases}$$

- active : $\partial L/\partial \lambda = 2x + 2y - 1 \overset{!}{=} 0$ ③

  Solving this system of three equations (①,②,③)

  gives $x = y = 0.25$. $\lambda = 0.125 > 0$. $\Rightarrow -x \cdot y = -\frac{1}{16}$

- inactive : $\lambda = 0 \Rightarrow x = y = 0$ ($\because$ ①,②)

  $\Rightarrow g(x, y) = -1 < 0$. $\Rightarrow -x \cdot y = 0$

  (o.k)

Q & A