Foundation of Machine Learning 5주차

정재헌, 우지수/23.02.17

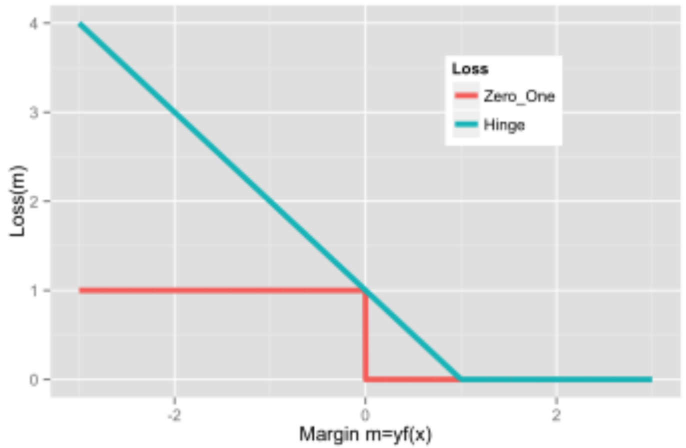
Computational Data Science LAB

CONTENTS

Margin and hinge loss

- The Margin
  - ✓ 정의: 예측 score $\hat{y}$ 과 실제 class $y \in \{-1,1\}$에 대한 마진은 $y\hat{y}$ 이다.
  - ✓ $yf(x)$, $f(x)$ 는 score function
  - ✓ 대부분 classification loss는 마진에만 의존

- Hinge Loss
  - ✓ SVM/Hinge loss: $\ell_{Hinge} = \max\{1-m, 0\} = (1-m)_+$
  - ✓ 마진 $m = yf(x)$; "positive part"인 $(x)_+ = x1(x \geq 0)$
  - ✓ m=1일 때 미분이 불가
  - ✓ Margin error



Support Vector Machine

- Hypothesis space
  - ✓ $\mathcal{F} = \{f(x) = w^T\, x + b \mid w \in R^d, b \in R\}$
- SVM Optimization Problem
  - ✓ $\min_{w \in R^d, b \in R} \frac{1}{2} \left||w|\right|^2 + \frac{c}{n} \sum_{i=1}^{n} \max(0, 1 - y_i\, [w^T\, x_i + b])$
  - ✓ Max가 있기에 미분이 불가
    - $minimize \quad \frac{1}{2}\left||w|\right|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i$
    - $subject\ to \quad \xi_i \geq \max(0, 1 - y_i[w^T x_i + b])$

- $minimize \quad \frac{1}{2}\left||w|\right|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i$
- $subject\ to \quad \xi_i \geq (1 - y_i[w^T x_i + b]) \quad for\ \ i = 1, \ldots, n$
- $\xi_i \geq 0 \quad for\ \ i = 1, \ldots, n$

- SVM Optimization Problem
  - ✓ 최종적으로 다음과 같은 식 도출
    - $minimize \quad \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i$
    - $subject\ to \quad (1 - y_i[w^T x_i + b]) - \xi_i \le 0 \quad for\ \ i = 1, \dots, n$
    - $\qquad\qquad -\xi_i \le 0 \quad for\ \ i = 1, \dots, n$
  - ✓ 위 식은 objective function이 이차식이고 제약식이 모두 affine이기에 quadratic program
  - ✓ 미분이 가능

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

- SVM Lagrange Multipliers
  - $L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}\left||w|\right|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i[w^T x_i + b] - \xi_i) + \sum_{i=1}^{n} \lambda_i(-\xi_i)$
  - $= \frac{1}{2}w^T w + \sum_{i=1}^{n} \xi_i(\frac{c}{n} - \alpha_i - \lambda_i) + \sum_{i=1}^{n} \alpha_i(1 - y_i[w^T x_i + b])$
  - $p^* = \inf_{w,\xi,b} \sup_{\alpha,\lambda \geqslant 0} L(w, b, \xi, \alpha, \lambda) \geq \sup_{\alpha,\lambda \geqslant 0} \inf_{w,\xi,b} L(w, b, \xi, \alpha, \lambda) = d^*$

✓ Slater’s condition(strong duality 만족)

✓ 따라서, SVM을 Quadratic Program으로 해결이 가능

- SVM Dual Function
  - $g(\alpha, \lambda) = \inf_{w,b,\xi} L(w, b, \xi, \alpha, \lambda) = \inf_{w,b,\xi} [\frac{1}{2} w^T w + \sum_{i=1}^{n} \xi_i (\frac{c}{n} - \alpha_i - \lambda_i) + \sum_{i=1}^{n} \alpha_i (1 - y_i [w^T x_i + b])]$

✓ Strong duality를 만족($p^* = d^*$)→ KKT condition 만족

✓ KKT condition – stationary condition

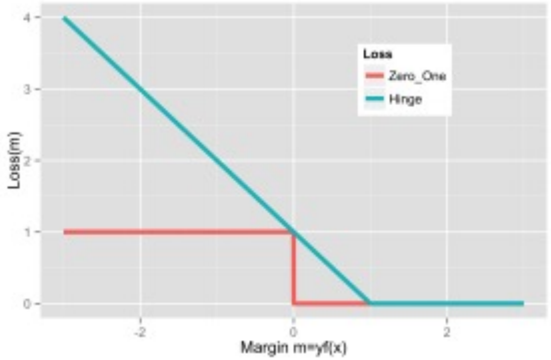
- SVM Dual Function: First Order Conditions
  - $\partial_w L = 0 \Leftrightarrow w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Leftrightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$
  - $\partial_b L = 0 \Leftrightarrow -\sum_{i=1}^{n} \alpha_i y_i = 0 \Leftrightarrow 0 = \sum_{i=1}^{n} \alpha_i y_i$
  - $\partial_{\xi_i} L = 0 \Leftrightarrow \frac{c}{n} - \alpha_i - \lambda_i = 0 \Leftrightarrow \frac{c}{n} = \alpha_i + \lambda_i$

- SVM Dual Function
  - ✓ First order condition에 의해서 라그랑지안 primal 문제가 Dual 문제로 바뀜
  - ✓ α_i에 관한 문제로 단순해짐(primal solution에서의 w를 구할 때, $w = \sum_{i=1}^{n} \alpha_i y_i x_i$이기에)
    - $\sup_{\alpha} \quad \sum_{i=0}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_i$
    - $s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0$
    - $\alpha_i \in \left[0, \frac{c}{n}\right], \quad i = 1, \dots, n$

02 | Complementary Slackness: Margin and Support Vectors

Insight From Complementary Slackness: Margin and Support Vectors

- The Margin
  - ✓ $f^*(x) = x^T w^* + b^*$이 존재한다고 가정할 때 마진 $yf^*(x)$는 다음과 같음



- Support Vectors and The Margin
  - ✓ "Slack Variable"이라고 부르는 $\xi_i^* = \max(0, 1 - y_i f^*(x_i))$는 (x_i, y_i)에서의 hinge loss
  - ✓ $\xi_i^* = 0$이라고 가정하면 그 때의 $y_i f^*(x_i) \geq 1$
  - ✓ 분류가 올바르게 된 $y_i f^*(x_i) \geq 1$의 hinge loss 0

Complementary Slackness

- Complementary Slackness Conditions
  - ✓ First order condition에서의 $\nabla_{\xi_i} L = 0$은 $\lambda_i^* = \frac{c}{n} - \alpha_i^*$를 도출
  - ✓ Strong duality(KKT condition)에 의해서, complementary slackness를 가짐
    - $\alpha_i^*(1 - y_i f^*(x_i) - \xi_i^*) = 0$
    - $\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$(1 - y_i f(x_i)) - \xi_i \leqslant 0$

- Consequences of Complementary Slackness Conditions
  - ✓ 만약 $y_i f^*(x_i) > 1$이면 그 때의 margin loss는 $\xi_i^* = 0$이되고 위의 식에 대입해본다면 $\alpha_i^* = 0$
  - ✓ 만약 $y_i f^*(x_i) < 1$이면 그 때의 margin loss는 $\xi_i^* > 0$이되고 위의 식에 대입해본다면 $\alpha_i^* = \frac{c}{n}$
  - ✓ 만약 $\alpha_i^* = 0$이면 그 때의 margin loss는 $\xi_i^* = 0$이되고 이 말인 즉슨 loss가 없다는 것을 의미해 최종적으로 $y_i f^*(x_i) \geq 1$
  - ✓ 만약 $\alpha_i^* \in \left(0, \frac{c}{n}\right)$이면 그 때의 margin loss는 $\xi_i^* = 0$이되고 $1 - y_i f^*(x_i) = 0$.

02 | Complementary Slackness: Margin and Support Vectors

Consequences of Complementary Slackness Conditions

- $\alpha_i^* = 0 \Rightarrow \; y_i f^*(x_i) \geq 1$
- $\alpha_i^* \in \left(0, \frac{c}{n}\right) \Rightarrow \; y_i f^*(x_i) = 1$
- $\alpha_i^* = \frac{c}{n} \Rightarrow \; y_i f^*(x_i) \leq 1$

- $y_i f^*(x_i) < 1 \Rightarrow \; \alpha_i^* = \frac{c}{n}$
- $y_i f^*(x_i) = 1 \Rightarrow \; \alpha_i^* \in \left[0, \frac{c}{n}\right]$
- $y_i f^*(x_i) > 1 \Rightarrow \; \alpha_i^* = 0$

Support Vector

- Support Vector
  - ✓ 만약 α^*가 dual problem의 해라고 하면, primal의 해는 다음과 같음
  - ✓ $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \quad , with \ \alpha_i^* \in \left[0, \frac{c}{n}\right]$
- The Bias Term: b
  - ✓ SVM primal에서 complementary slackness condition
    - $\alpha_i^*(1 - y_i f^*(x_i) - \xi_i^*) = 0$ (1)
    - $\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$ (2)

- ✓ $\alpha_i^* \in \left(0, \frac{c}{n}\right)$와 같은 i가 있다고 가정
- ✓ (2)에 의해서 $\xi_i^* = 0$을 만족
- ✓ (1)에 의해서 다음식을 만족
  - $y_i\left[x_i^T w^* + b^*\right] = 1$
  - $\Leftrightarrow \left[x_i^T w^* + b^*\right] = y_i \quad (use\ y_i \in \{-1,1\})$
  - $\Leftrightarrow b^* = y_i - x_i^T w^*$

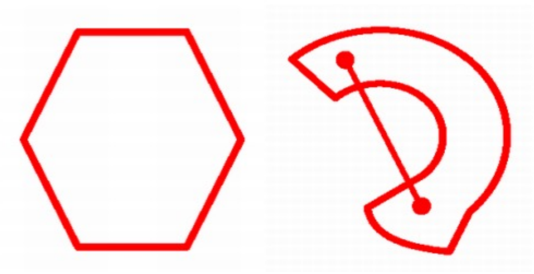
Support Vector

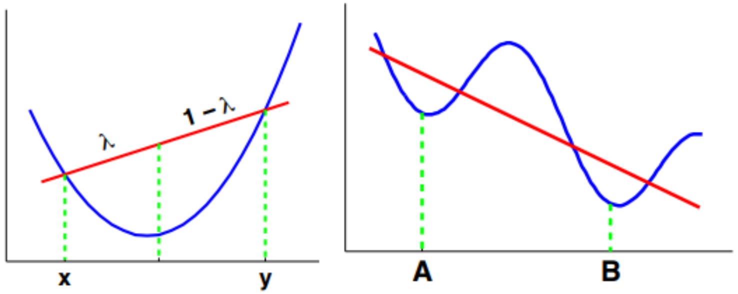
- Support Vector
  - ✓ 만약 α^*가 dual problem의 해라고 하면, primal의 해는 다음과 같음
  - ✓ $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \quad , with\ \ \alpha_i^* \in \left[0, \frac{c}{n}\right]$
- The Bias Term: b
  - ✓ SVM primal에서 complementary slackness condition
    - $\alpha_i^*(1 - y_i f^*(x_i) - \xi_i^*) = 0$ (1)
    - $\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$ (2)

- ✓ $\alpha_i^* \in \left(0, \frac{c}{n}\right)$와 같은 i가 있다고 가정
- ✓ (2)에 의해서 $\xi_i^* = 0$을 만족
- ✓ (1)에 의해서 다음식을 만족
  - $y_i\left[x_i^T w^* + b^*\right] = 1$
  - $\Leftrightarrow \left[x_i^T w^* + b^*\right] = y_i \qquad (use\ y_i \in \{-1,1\})$
  - $\Leftrightarrow b^* = y_i - x_i^T w^*$

03 | Subgradient
REVIEW

- Support Vector
  - ✓ SVM의 목적은 마진을 최대화하는 것
  - ✓ SVM에서는 비용 함수가 다른 대부분의 머신 러닝 알고리즘들과 달리 매끄럽지 않은 형태(미분불가)
  - ✓ 경사하강법을 이용한 알고리즘이 아닌 다른 형태의 알고리즘
- Convex Sets
  - ✓ 어떤 집합이 주어져 있다고 할 때, 이 집합의 원소인 두 점을 잇는 선분이 이 집합이 포함 될 때 우리는 이 집합을 convex set이라 함
  - ✓ 두 점을 잇는 선분이 그래프 위에 있다면 convex하다고 하고 –f가 convex하다면 f는 concave하다고 함





- Convex Optimization Problem: Standard Form
    - $minimize \quad f_0(x)$
    - $subject\ to \quad f_i(x) \leq 0, \quad i = 1, \ldots, m$

✓ 이 때, $f_0, \ldots, f_m$은 모두 convex 함수

- Level Sets and Sublevel Sets

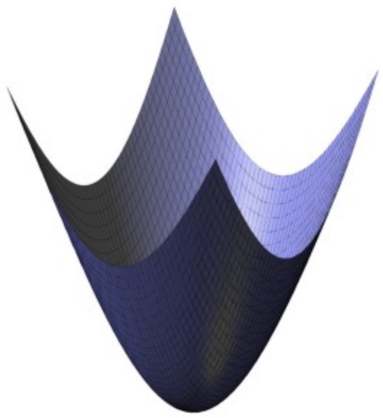
✓ $f: \mathbb{R}^d \rightarrow \mathbb{R}$ 인 함수

✓ $f(x) = c$에 대해서 $x \in \mathbb{R}^d$에 모든 점들의 집합을 value c에 대한 level set 혹은 contour line이라 함

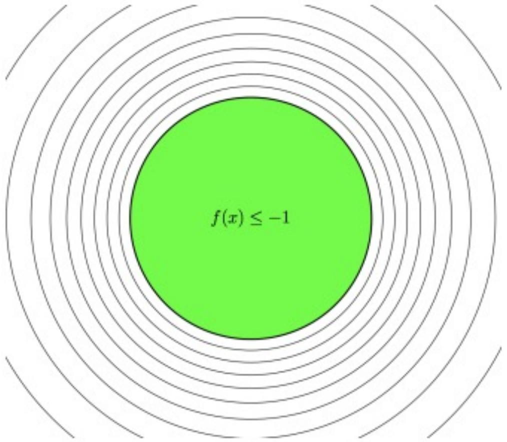
✓ $f(x) \leq c$에 대해서 $x \in \mathbb{R}^d$에 모든 점들의 집합을 value c에 대한 sublevel set이라 함

✓ $f: \mathbb{R}^d \rightarrow \mathbb{R}$함수가 convex하다면 그 때의 sublevel set도 convex하다고 함

Level Sets and Sublevel Sets : convex

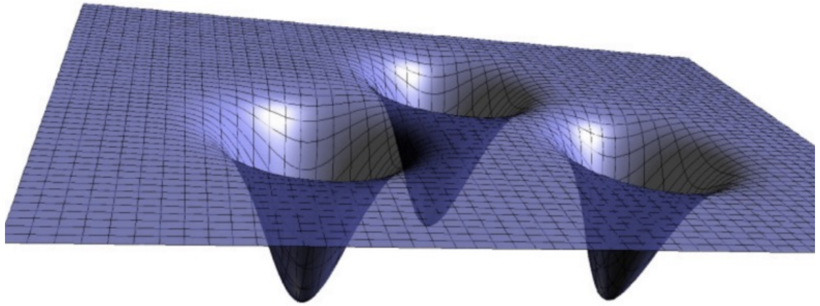


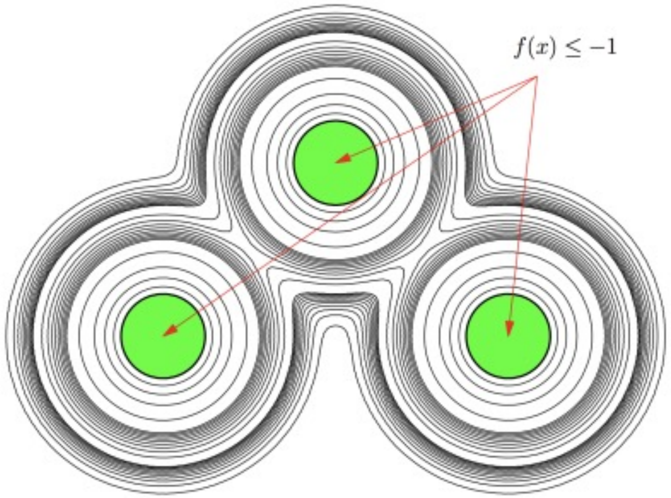
- Level Sets and Sublevel Sets
  - ✓ 이러한 convex function이 있다고 하자, 이때 floor는 domain으로서 2차원이고 그래프는 3차원
  - ✓ 중간에 평면으로 자른 지점이 level set이고 그 아랫부분이 sublevel set



- Level Sets and Sublevel Sets(contour plot)
  - ✓ 그리고 이걸 단면으로 봤을 때 중간에 있는 부분이 sublevel set이라고 할 수 있음

Level Sets and Sublevel Sets : non-convex

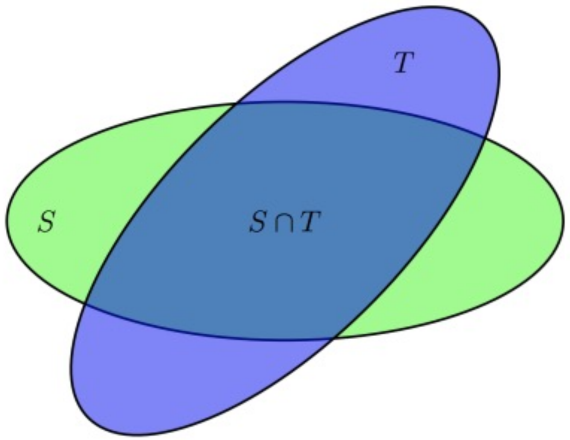


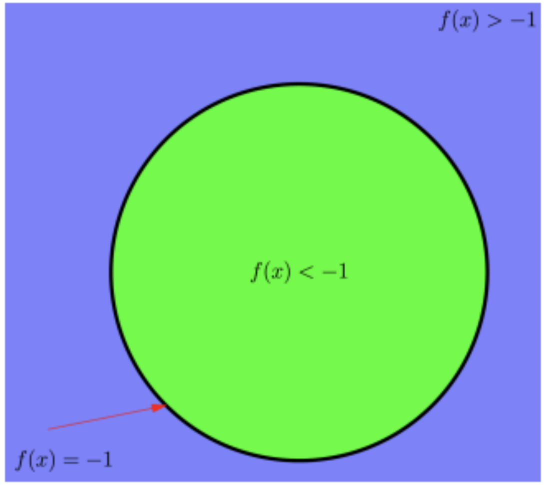


- Level Sets and Sublevel Sets
  - ✓ 3차원 그래프에서 non-convex함을 볼 수 있음

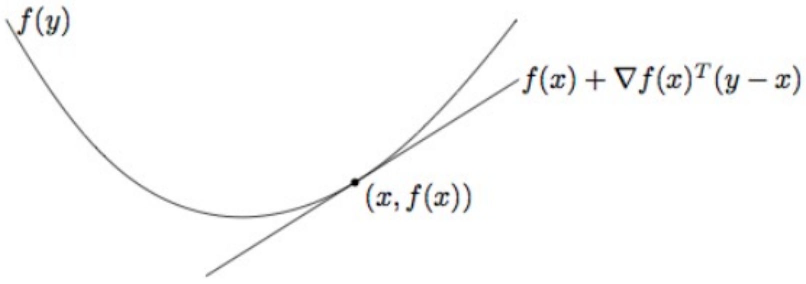
- Level Sets and Sublevel Sets(contour plot)
  - ✓ contour plot으로 봤을 때 sublevel set에서 중간점을 이어보면 non-convex함을 알 수 있음

- Level Sets and Sublevel Sets
  - ✓ Convex set의 특징으로는 convex set들의 교집합 또한 convex set 이라는 성질을 가진다는 점
  - ✓ 또 다른 특징으로는 그냥 level set이나 superlevel set(f(x)>-1인 부분)은 일반적으로 convex set이 아닐 수도 있다는 특징이 있음
  - ✓ 결과적으로 C를 convex set이라고 했을 때 우리는 Convex Optimization Problem을 다음과 같이 쓸 수 있음
    - $minimize \quad f_0(x)$
    - $subject\ to \quad x \in C$
  - ✓ 이때 f는 convex 함수이고 C는 convex set



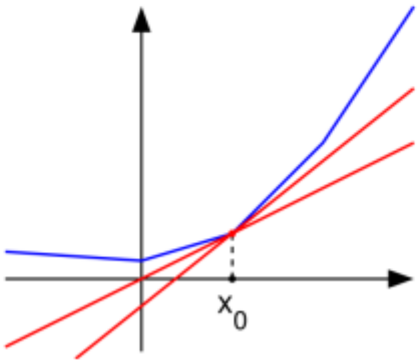


- First-Order Approximation(1차 근사)
  - ✓ $f: \mathbb{R}^d \rightarrow \mathbb{R}$가 미분 가능하다고 가정
  - ✓ y를 추정하기 위해 1차 선형근사를 시키는 것을 알 수 있음
  - ✓ $f(y) \approx f(x) + \nabla f(x)^T(y - x)$
  - ✓ 만약 함수 f 가 미분 가능하고, convex function 이라면 이것을 global underestimator이라 함
  - ✓ $f(y) \geq f(x) + \nabla f(x)^T(y - x)$
  - ✓ $\nabla f(x) = 0$이라면 그 때 x는 f의 global minimizer



Subgradient

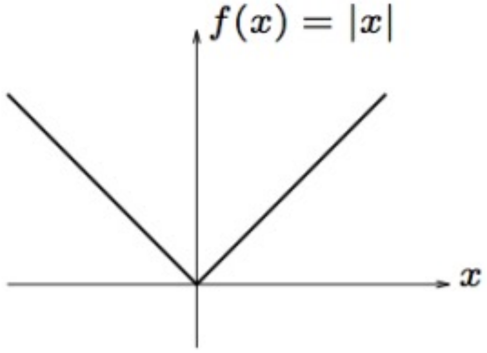
- Subgradients
  - ✓ $f(z) \geq f(x) + g^T(z - x)$
  - ✓ 이 때 g를 subgradient라고 함
  - ✓ 이 때 빨간선들은 앞서 살펴봤던 global underestimator이고 그 안에 g가 subgradient
- Subdifferential
  - ✓ 만약 x에서 subgradient가 적어도 하나라도 존재한다면 f는 x에서의 subdifferentiable
  - ✓ x에서의 모든 subgradient의 집합은 $\partial f(x)$인 subdifferential이라고 부름
  - ✓ 만약 f가 convex하고 미분 가능하다면 → $\partial f(x) = \{\nabla f(x)\}$
  - ✓ 만약 $\partial f(x) = \emptyset$ → f는 convex하지 않다

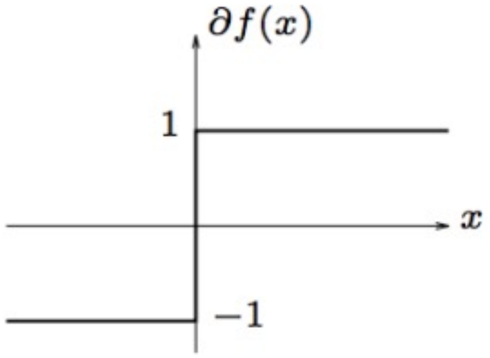


- Global Optimality Condition
  - ✓ vector $g \in \mathbb{R}^d$는 모든 z에 대해 x에서 $f: \mathbb{R}^d \to \mathbb{R}$의 subgradient
  - ✓ $f(z) \geq f(x) + g^T(z - x)$
  - ✓ 만약 $0 \in \partial f(x)$이면 그때의 x는 f의 global minimizer

Subdifferential of Absolute Value

- Subdifferential of Absolute Value
  - ✓ $f(x) = |x|$
  - ✓ $\{(x, g) | x \in \mathbb{R}, g \in \partial f(x)\}$

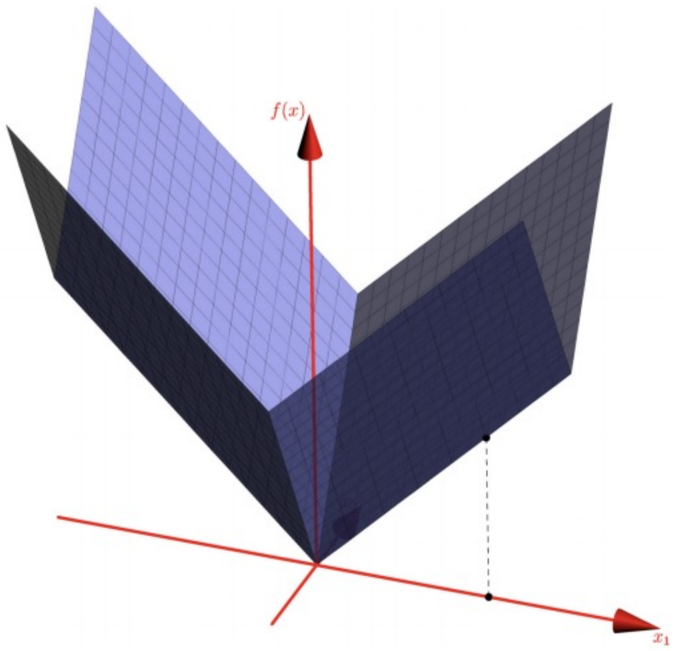




03 | Subgradient

$$f(x_1, x_2) = |x_1| + 2|x_2|$$

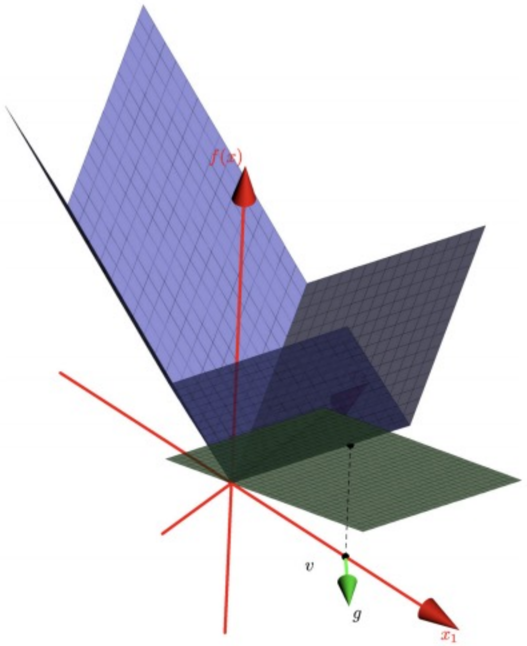
- Subgradients of $f(x_1, x_2) = |x_1| + 2|x_2|$
  - ✓ $f(x_1, x_2) = |x_1| + 2|x_2|$에서 (3, 0)지점의 subdifferential
  - ✓ $|x_1|$의 subgradient는 1($x_1 = 3$)
  - ✓ $|x_2|$의 subgradient는 [-2,2]($x_2 = 0$)
  - ✓ $h(x_1, x_2) = f(3,0) + g^T(x_1 - 3, x_2 - 0)$
  - ✓ 이때, h는 초평면이자 $f(x_1, x_2)$의 global underestimate
  - ✓ $g = (g_1, g_2)$
  - ✓ $g_1 = 1, \quad g_2 = [-2,2]$



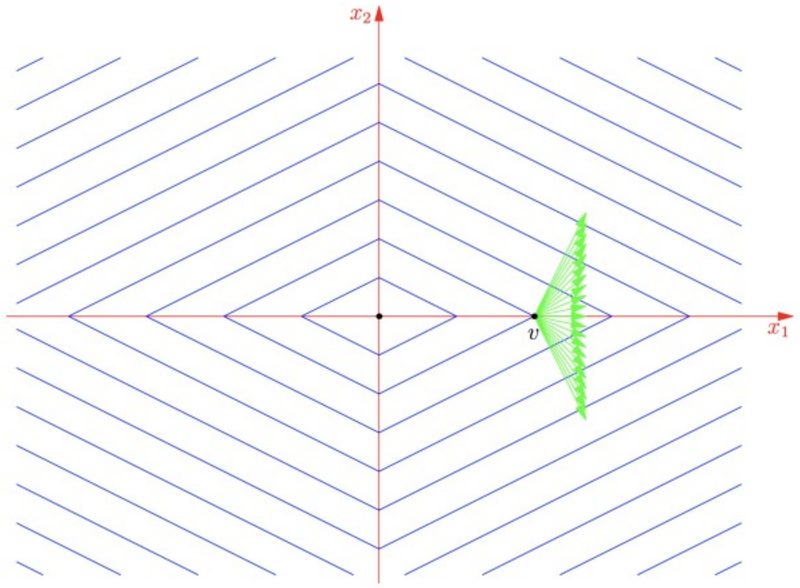
03 | Subgradient

$f(x_1, x_2) = |x_1| + 2|x_2|$

- Underestimating Hyperplane to $f(x_1, x_2) = |x_1| + 2|x_2|$
  - ✓ $f(x_1, x_2) = |x_1| + 2|x_2|$에서 (3, 0)지점의 subdifferential
  - ✓ $|x_1|$의 subgradient는 1$(x_1 = 3)$
  - ✓ $|x_2|$의 subgradient는 [-2,2]$(x_2 = 0)$
  - ✓ $h(x_1, x_2) = f(3,0) + g^T(x_1 - 3, x_2 - 0)$
  - ✓ 이때, h는 초평면이자 $f(x_1, x_2)$의 global underestimate
  - ✓ $g = (g_1, g_2)$
  - ✓ $g_1 = 1, \quad g_2 = [-2,2]$

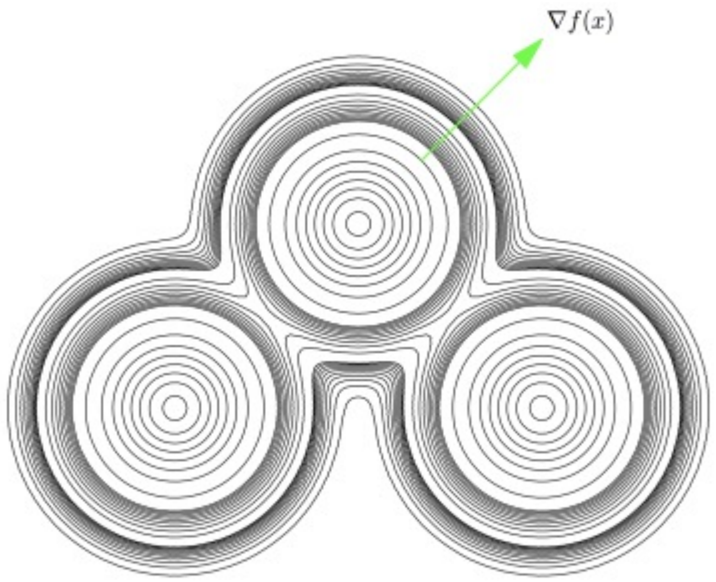


- Subdifferential on Contour Plot

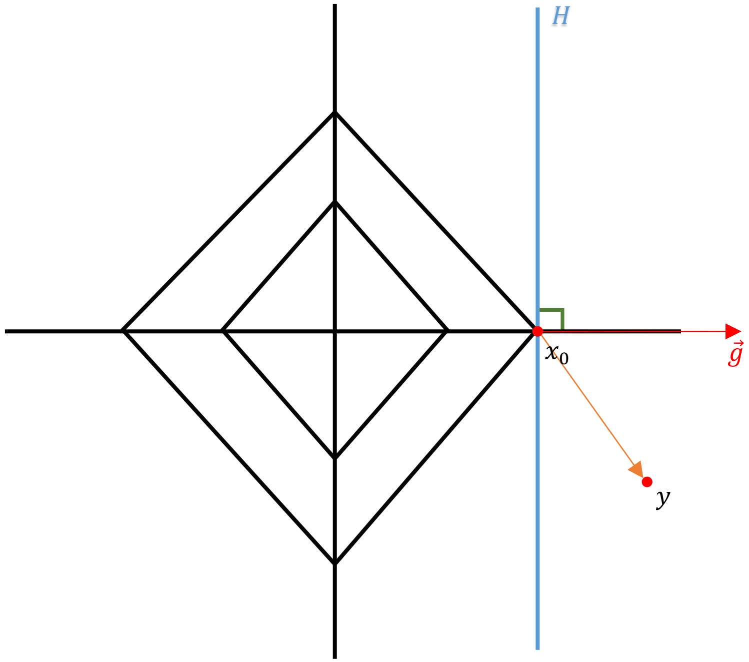


Contour Line and Gradient

- Contour Line and Gradient
  - ✓ $f: \mathbb{R}^d \rightarrow \mathbb{R}$에서 함수의 그래프는 $\mathbb{R}^{d+1}$차원에서 존재
  - ✓ f의 gradient나 subgradient는 $\mathbb{R}^d$차원에 존재
  - ✓ 또한 등고선, level set그리고 sublevel set모두 $\mathbb{R}^d$에 존재
  - ✓ $f: \mathbb{R}^d \rightarrow \mathbb{R}$이 연속적으로 미분이 가능하고 $\nabla f(x_0) \neq 0$
  - ✓ 그 때 $\nabla f(x_0)$은 level set S에 대해서 수직이라고 할 때
  - ✓ Level sets: $S = \{x \in \mathbb{R}^d | f(x) = f(x_0)\}$

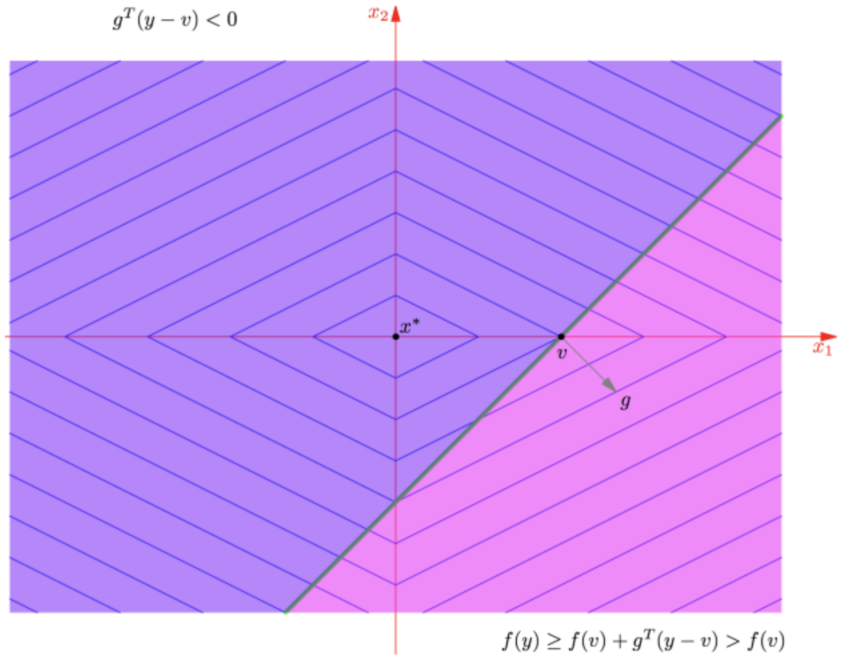


- Contour Line and Subgradients
  - ✓ $f : \mathbb{R}^d \to \mathbb{R}$, 점 x_0에서 subgradient g가 존재
  - ✓ g에 직교하는 초평면 H level set $S = \{x \in \mathbb{R}^d | f(x) = f(x_0)\}$를 support
    (S에서 모든 점들이 H의 한쪽에 위치한다는 뜻)
- Proof:
  - ✓ 임의의 y에서 $f(y) \geq f(x_0) + g^T(y - x_0)$
  - ✓ 만약 y가 g가 양수인 방향에 위치한다면, $y - x_0$는 g와 평행하지 않으며 두 vector 의 내역은 0보다 큼
  - ✓ 그렇다면 $f(y)$는 $f(x_0)$의 값보다 커짐, 즉 y는 S에 포함되지 않음
  - ✓ 따라서 모든 S의 원소들은 H위나 H의 –g 방향에 존재



Subgradient of $f(x_1, x_2) = |x_1| + 2|x_2|$

- Subgradient of $f(x_1, x_2) = |x_1| + 2|x_2|$
  - ✓ 증명에 의해 g가 가리키는 방향에 위치한 점들은 함수 f보다 큰 값을 가짐
  - ✓ 즉, 이 점들은 함수 f를 더 크게 만드는 방향
  - ✓ 반면 g의 반대 방향은 함수 f보다 작을수도 있고 같을 수도 있음
  - ✓ 이러한 이유로 항상 –g가 감소하는 방향이 아닐 수 있음



Subgradient Descent

- Subgradient Descent
  - ✓ f가 convex하고 x_0가 최적화 시작점이라 가정
  - ✓ 음수의 subgradient 방향에서의 step: $x = x_0 - tg, \quad t > 0 \; and \; g \in \partial f(x_0)$
- Subgradient Gets Us Closer To Minimizer
  - ✓ z가 임의의 점에서 $f(z) < f(x_0)$을 만족
  - ✓ 그렇다면 충분히 작은 $t > 0$에 대해서 $\| x - z \|_2 < \| x_0 - z \|_2$ 또한 만족
  - ✓ $z = x^* \in \operatorname*{argmin}_x f(x)$로 적용한다면 음수의 subgradient 방향에서의 step은 최소화에 더 가깝게 만들어줌

Subgradient Gets Us Closer To Minimizer

- Subgradient Gets Us Closer To Minimizer(Proof)
  - ✓ $\| x - z \|_2^2 = \| x_0 - tg - z \|_2^2$
  - ✓ $= \| x_0 - z \|_2^2 - 2\,\mathrm{tg}^T(x_0 - z) + t^2 \| g \|_2^2$
  - ✓ $\leq \| x_0 - z \|_2^2 - 2\,t[f(x_0) - f(z)] + t^2 \| g \|_2^2$

- $-2\,t[f(x_0) - f(z)] + t^2 \| g \|_2^2$
  - ✓ Convex quadratic 문제이다.
  - ✓ $t = 0$이고 $t = 2\big(f(x_0) - f(z)\big)/\| g \|_2^2 > 0$일 때 0을 갖게 됨
  - ✓ 그러므로, 이는 임의의 $t \in \left(0, \frac{2(f(x_0) - f(z))}{\|g\|_2^2}\right)$에서 음수

04 | Subgradient Descent

Convergence Theorem

- ✓ $f: \mathbb{R}^d \to \mathbb{R}$가 convex하고, f가 $G > 0$제약조건을 가진 Lipschiz 연속일 때
- ✓ $| f(x) - f(y) | \leq G \| x - y \|, \quad for\ all\ x, y$

- Fixed Step Size
  - ✓ 고정된 step size t에서, $\lim_{k\to\infty} f(x_{best}^{k}) \leq f(x^*) + G^2\frac{t}{2}$

- Decreasing Step Sizes
  - ✓ 로빈스-몬로 조건과 관련된 step size의 경우, $\lim_{k\to\infty} f(x_{best}^{k}) \leq f(x^*)$

Q&A

감사합니다.