Subgradients

Ryan Tibshirani
Convex Optimization 10-725

Last time: gradient descent

Consider the problem

$$\min_x \; f(x)$$

for f convex and differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$. Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Step sizes t_k chosen to be fixed and small, or by backtracking line search

If ∇f is Lipschitz, gradient descent has convergence rate $O(1/\epsilon)$. Downsides:

- Requires f differentiable
- Can be slow to converge

· Review.

Gradient Descent : $x^{(k)} = x^{(k-1)} - t\nabla f(x)$

f : Diff, Convex, and Lipsitz conti with L, then for $t < \frac{1}{L}$.

$$|f(x^{(k)}) - f^*| < \frac{1}{2tk}\|x^{(0)} - x^*\|_2^2$$

error : $O(\frac{1}{k})$, step : $O(\frac{1}{\epsilon})$ $\rightsquigarrow \frac{1}{0.0001}$: 10000 스텝

G.D with Strong Convexity

f : Diff, Strong Convex with m and Lipsitz conti with L, then for $t < \frac{2}{L+m}$

$$|f(x^{(k)}) - f^* < \gamma^k L\|x^{(0)} - x^*\|_2^2$$

error : $O(e^{-k})$, step : $O(\log 1/\epsilon)$ $\rightsquigarrow \log\frac{1}{0.0001}$: 4스텝

$$mI < \nabla^2 f(x) < LI.$$

$\Rightarrow \lambda_1 < x^t\nabla^2 f(x)x < \lambda_n$

· but it assume not ~~differentiable of~~ f.

문제 상황

⇒ Cannot use First Order Condition (Gradient Descent)

Outline

Today: crucial mathematical underpinnings!

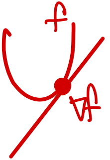
- Subgradients
- Examples
- Properties
- Optimality characterizations

Subgradients : 새로운 컨셉

: 미분이 불가능하니까 대신 사용.

Recall that for convex and differentiable f,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y$$

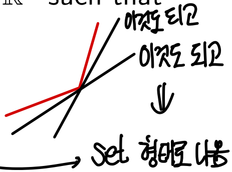


That is, linear approximation always underestimates f

A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

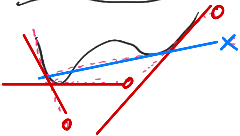
$$f(y) \geq f(x) + \boxed{g^T} (y - x) \quad \text{for all } y$$

∇f(x)를 대신하는 "기울기"



- Always exists[1]
- If f differentiable at x, then $g = \nabla f(x)$ uniquely
- Same definition works for nonconvex f (however, subgradients need not exist)

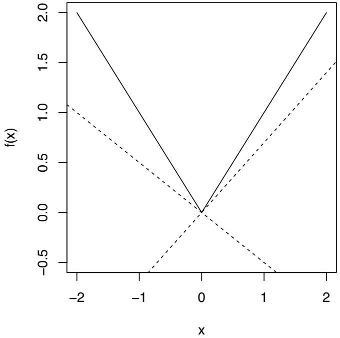
Convex 아니어도
subgradient가 성립한다!!



[1]On the relative interior of $\mathrm{dom}(f)$

Examples of subgradients

Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$

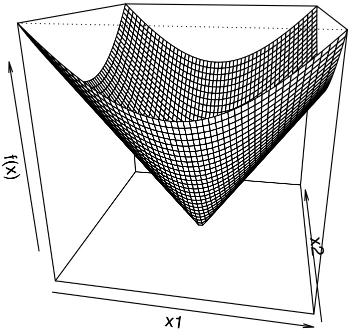


- For $x \neq 0$, unique subgradient $g = \operatorname{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2 = \sqrt{x^T x}$

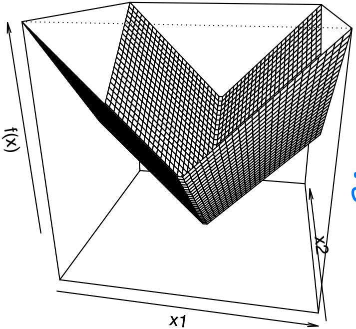
$x^T x$는 항상 미분 O,

$\sqrt{x^T x}$는 $x=0$에서 안됨.



- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$



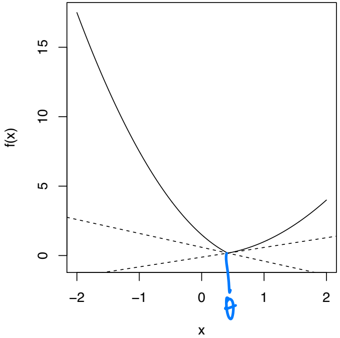
$x = (3, -1)$

$\partial f(x) = (1, -1)$

- For $x_i \neq 0$, unique ith component $g_i = \text{sign}(x_i)$
- For $x_i = 0$, ith component g_i is any element of $[-1, 1]$

Consider $f(x) = \max\{f_1(x), f_2(x)\}$, for $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ convex, differentiable

미분X,



- For $f_1(x) > f_2(x)$, unique subgradient $g = \nabla f_1(x)$
- For $f_2(x) > f_1(x)$, unique subgradient $g = \nabla f_2(x)$
- For $\boxed{f_1(x) = f_2(x),}$ subgradient g is any point on line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$

$[\nabla f_1(\theta), \nabla f_2(\theta)].$

$\nabla f(x)$: f의 gradient : nabla

$\partial f(x)$: f의 subgradient
여럿 할 수 있는 집합 : partial.

Subdifferential

Subgradient 집합

Set of all subgradients of convex f (on x.) is called the subdifferential:

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$$

- Nonempty (only for convex f)
- $\partial f(x)$ is closed and convex (even for nonconvex f)
- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$ (원소가 오직 하나)
- If $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$
  하나이면　　미분가능.

Connection to convex geometry

Convex set $C \subseteq \mathbb{R}^n$, consider indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$,

min $f(x) + I_C(x)$

$$I_C(x) = I\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

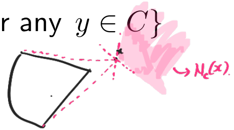
C에 안들어가면 → ∞ 이니까 not feasible

For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x is, recall

$$\mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

$0 \leq g^T(x-y) = \|g\| \|x-y\| \cos\theta$

g와 $(x-y)$의 각도가 $90°$ 이상

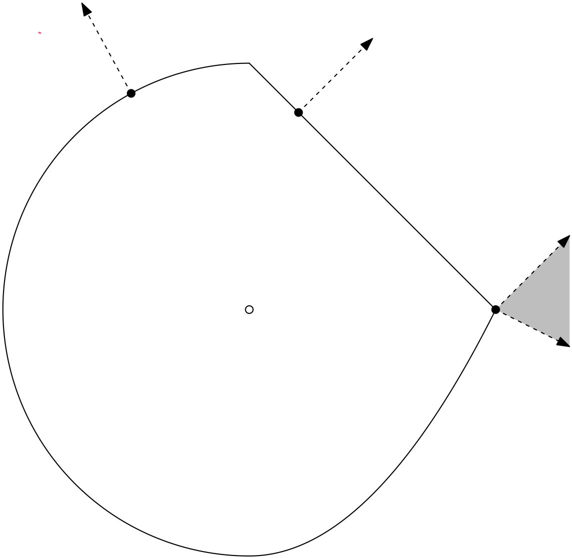


Why? By definition of subgradient g,

$(x-y)$

$$I_C(y) \geq I_C(x) + g^T(y - x) \quad \text{for all } y$$

- For $y \notin C$, $I_C(y) = \infty$
- For $y \in C$, this means $0 \geq g^T(y - x)$



Subgradient calculus

subgradient 계산에 쓰이는 기법·규칙.

Basic rules for convex functions:

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- Affine composition: if $g(x) = f(Ax + b)$, then

$$\partial g(x) = A^T \partial f(Ax + b)$$

- Finite pointwise maximum: if $f(x) = \max_{i=1,\ldots,m} f_i(x)$, then

$$\partial f(x) = \underset{\text{convex hull}}{\text{conv}} \left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

convex hull of union of subdifferentials of active functions at x

$\partial f_1(a) = \{1\}$

$\partial f_2(a) = \{-1\}$

$\partial f_1(a) \cup \partial f_2(a) = \{-1, 1\}$ $\quad = \{-1 \le x \le 1\}$

$\text{Conv}(\partial f_1(a) \cup \partial f_2(a)) = \text{Conv}(\{-1, 1\}) = \{x ; -1 \cdot t + 1 \cdot (1-t), t \in [0,1]\}$

- General composition: if

$$f(x) = h\big(g(x)\big) = h\big(g_1(x), \ldots, g_k(x)\big)$$

where $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^k \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$, h is convex and nondecreasing in each argument, g is convex, then

$$\partial f(x) \subseteq \Big\{ p_1 q_1 + \cdots + p_k q_k : \\ p \in \partial h(g(x)),\ q_i \in \partial g_i(x),\ i = 1, \ldots, k \Big\}$$

- General pointwise maximum: if $f(x) = \max_{s \in S} f_s(x)$, then

$$\partial f(x) \supseteq \mathrm{cl} \Bigg\{ \mathrm{conv} \Bigg(\bigcup_{s : f_s(x) = f(x)} \partial f_s(x) \Bigg) \Bigg\}$$

Under some regularity conditions (on S, f_s), we get equality

- Norms: important special case. To each norm $\|\cdot\|$, there is a dual norm $\|\cdot\|_*$ such that

$$\|x\| = \max_{\|z\|_* \le 1} z^T x$$

holder's inequality

(For example, $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual when $1/p + 1/q = 1$.) In fact, for $f(x) = \|x\|$ (and $f_z(x) = z^T x$), we get equality:

$$\partial f(x) = \mathrm{cl}\left\{\mathrm{conv}\left(\bigcup_{z: f_z(x) = f(x)} \partial f_z(x)\right)\right\}$$

Note that $\partial f_z(x) = z$. And if z_1, z_2 each achieve the max at x, which means that $z_1^T x = z_2^T x = \|x\|$, then by linearity, so will $tz_1 + (1-t)z_2$ for any $t \in [0,1]$. Thus

$$\partial f(x) = \operatorname*{argmax}_{\|z\|_* \le 1} z^T x$$

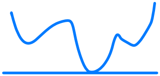
Optimality condition

Recall. first order optimality condition

f: diff & convex $\Rightarrow f(x^*) \geq \nabla f(y)(x^* - y), \forall y$

For any f (convex or not),

$$f(x^\star) = \min_x f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$



That is, $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$. This is called the **subgradient optimality condition**

Why? Easy: $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note the implication for a convex and differentiable function f, with $\partial f(x) = \{\nabla f(x)\}$

convex & diff 함수 f에 대해 $f'=0$ 해서 찾았는데, subgradient 하면 non-convex not diff 에서도 가능.

Derivation of first-order optimality

Example of the power of subgradients: we can use what we have learned so far to derive the first-order optimality condition. Recall

$$\min_x \; f(x) \;\; \text{subject to} \;\; x \in C$$

is solved at x, for f convex and differentiable, if and only if

$$\nabla f(x)^T (y - x) \geq 0 \;\; \text{for all} \;\; y \in C$$

Intuitively: says that gradient increases as we move away from x. How to prove it? First recast problem as

$$\min_x \; f(x) + I_C(x)$$

Now apply subgradient optimality: $0 \in \partial(f(x) + I_C(x))$

Observe

$$
\begin{aligned}
0 \in \partial\big(f(x) + I_C(x)\big) \\
&\iff \quad 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x) \\
&\iff \quad -\nabla f(x) \in \mathcal{N}_C(x) \\
&\iff \quad -\nabla f(x)^T x \geq -\nabla f(x)^T y \text{ for all } y \in C \\
&\iff \quad \nabla f(x)^T (y - x) \geq 0 \text{ for all } y \in C
\end{aligned}
$$

as desired

Note: the condition $0 \in \partial f(x) + \mathcal{N}_C(x)$ is a fully general condition for optimality in convex problems. But it's not always easy to work with (KKT conditions, later, are easier)

Example: lasso optimality conditions

단순하지만 강력하다...!

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, lasso problem can be parametrized as

$$\min_{\beta} \; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda \geq 0$. Subgradient optimality: $= \frac{1}{2}\partial\|y - X\beta\|_2^2 + \lambda\,\partial\|\beta_1\|_1$

$= -X^T(y - X\beta) + \lambda\,\partial\|\beta\|_1$

$$0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big)$$

$$\iff \quad 0 \in -X^T(y - X\beta) + \lambda\partial\|\beta\|_1$$

$$\iff \quad X^T(y - X\beta) = \lambda v$$

for some $v \in \partial\|\beta\|_1$, i.e.,

$\beta = (-4, 3, 2, 0)$

$\Downarrow$

$(-1, 1, 1, [-1, 1])$

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \ldots, p$$

$\frac{\partial}{\partial x} f(g(x)) = f'(g(x)) \cdot g'(x)$

예시, $f(x) = x^T x$, $\to f'(x) = 2x$

$g(\beta) = y - X\beta$. $\to g'(\beta) = -X$

$f(g(\beta)) = (y - X\beta)^T(y - X\beta)$

$\frac{\partial}{\partial x} f(g(\beta)) = f'(g(\beta))\, g'(\beta)$

$=$

Penalized Regression.

- (LSE + 페널티)를 통해 $\boxed{\text{다중공선성/오버피팅을 억제}}$하는 기법

벡터의 크기를 재는 단위 ex) $\begin{pmatrix}3\\4\end{pmatrix} \lessgtr \begin{pmatrix}2\\5\end{pmatrix} \Rightarrow \left\|\begin{pmatrix}3\\4\end{pmatrix}\right\|_2 = (3^2+4^2)^{1/2} = 5$

$\left\|\begin{pmatrix}2\\5\end{pmatrix}\right\|_2 = (2^2+5^2)^{1/2} = \sqrt{29}$

$$\min \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \|\beta\|_p^p \qquad \left(\text{(cf) } L_p\text{-Norm: } \|x\|_p = \left(\sum_{i=1}^{n} x_i^p\right)^{1/p},\ x \in \mathbb{R}^n \right)$$

$\|\cdot\|_p = (\Sigma(|x_i|)^p)^{1/p}$

일반적으로는 $\left\{ \begin{array}{lll} p=2 & \|\beta\|_2 = (\Sigma|\beta_i|^2)^{1/2} & (\text{aka. Ridge Regression}) \\ p=1 & \|\beta\|_1 = (\Sigma|\beta_i|) & (\text{aka. LASSO Regression}) \end{array} \right\}$을 사용.

- $\underbrace{\text{다중공선성}}_{\text{원인}}$ & $\underbrace{\text{오버피팅}}_{\text{원인.}}$으로 인한 $\underbrace{\text{모수 추정 오차}(Var(\hat{\beta}))}_{\text{결과.}}$을 잡기위한 알고리즘.

- 다중공선성 ⇒ X^tX 행렬이 singular 해지는 현상.

하다~ 가 아님.

$\left(\begin{array}{l} \Rightarrow \det(X^tX)=0. \\ \det(X^tX) = \prod_{i=1}^{n} \lambda_i(X^tX) \\ \quad \Rightarrow \text{아이겐밸류들 0이 존재하는 것.} \end{array} \right)$

(Pseudo-singular.
⇒ 아이겐밸류들 0에 매우 근접한 값이 존재하는 것)

X_1	=	$X_2+X_3+\varepsilon$	
4		1	3
3		2	1
2		1	1
5		4	1

왜 문제? $X^tX = T\Lambda T^t \Rightarrow (X^tX)^{-1} = T\Lambda^{-1}T^t$

$\Lambda^{-1} = \begin{pmatrix}\lambda_1^{-1} & & \\ & \lambda_2^{-1} & \\ & & \ddots \ \lambda_n^{-1}\end{pmatrix} \Rightarrow$ Singular 할시 $\lambda_n^{-1} = 0^{-1}$ 이므로. 아예 불가.

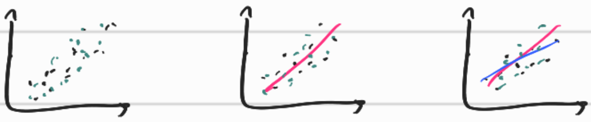
Pseudo singular 할시, λ_n^{-1}가 매우 큰 값이 되므로 $(X^tX)^{-1}$의 원소들이 매우 크게 됨.

λ_n^{-1}는 λ_n이 조금만 변해도 매우 크게 변하므로 $(X^tX)^{-1}$이 매우 불안정 해짐.
($Var((X^tX)^{-1})$가 커진다)

$\left(\text{ex) } \begin{array}{l} \lambda_n = 0.0001 \\ \lambda_{n-1} = 0.00005 \end{array} \Rightarrow \begin{array}{l} \lambda_n^{-1} = 10000. \\ \lambda_{n-1}^{-1} = 20000 \end{array} \right)$

오버피팅 $MSE(\hat{\beta}) = E(|Y - X\hat{\beta}|^2) = E(|Y - X\beta|^2) + E(|X\beta - X\hat{\beta}|^2)$

$\approx Bias(\beta) + \underline{Var(\hat{\beta})}\uparrow$



$\underset{\text{실제 데이터}}{(X, Y)} \Leftarrow \underset{\text{모델 (가설)}}{X\beta} \Leftarrow \underset{\text{구현된 모형}}{X\hat{\beta}}$

Bias는 가설이 실제를 얼마나 잘 표현할 수 있는지 : Variance는 가설을 얼마나 제대로 구현 할 수 있는지

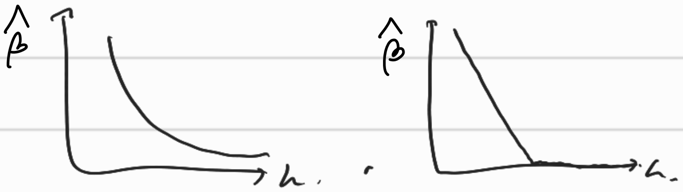
$$\min \ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_p^p$$

$\|Y-X\beta\|_2^2$ 를 minimize $\Rightarrow$ $\beta = (X^tX)^{-1}X^tY$.

$\|\beta\|_p$ 를 minimize $\Rightarrow$ $\beta = 0$.

즉 λ의 크기가 커짐에 따라 $\|\beta\|_p$의 중요도가 커지며

$\beta = (X^tX)^{-1}X^tY \longrightarrow 0$ 으로 바뀐다.



p=2   p=1

여기서 중요한 특징이,

- Ridge 의 경우 필요없는 β_i들 0에 근접하게 만들지만 완전히 0으로 만들진 못한다.
- LASSO 의 경우 필요없는 β_i들 아예 0으로 만듦.

실제 Solution을 살펴보자

- Ridge : $\min \ (Y-X\beta)^t(Y-X\beta) + \lambda\beta^t\beta$.

$$\nabla f(\beta) = X^t(Y-X\beta) + 2\lambda\beta$$
$$= X^tY - X^tX\beta + 2\lambda\beta$$
$$= X^tY - (X^tX+2\lambda I)\beta =: 0. \ (\text{F.O.C})$$

$$X^tY = (X^tX + 2\lambda I)\beta$$
$$\Rightarrow \beta^* = (X^tX + 2\lambda I)^{-1}X^tY.$$

(cf. 이를 Tikhonov-Regularized Inverse 라고 한다)

↳ 결국 reg의 closed-form solution

Write $X_1, \ldots, X_p$ for columns of X. Then our condition reads:

$$\begin{cases} X_i^T(y - X\beta) = \lambda \cdot \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda & \text{if } \beta_i = 0 \end{cases}$$

$\Rightarrow |X_i^T \hat{\varepsilon}| \Rightarrow X_i \perp \hat{\varepsilon}$.

Note: subgradient optimality conditions don't lead to closed-form expression for a lasso solution ... however they do provide a way to check lasso optimality

They are also helpful in understanding the lasso estimator; e.g., if $|X_i^T(y - X\beta)| < \lambda$, then $\beta_i = 0$ (used by screening rules, later?)

X_i가 필요가 없게.

$X_i \perp Y \Rightarrow Y$를 설명하는 X_i의 정보력이 없다.

Example: soft-thresholding

Simplfied lasso problem with $\overset{n\times p \,\to\, n\times n}{X = I}$:

$$\min_\beta \ \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|\beta\|_1$$

This we can solve directly using subgradient optimality. Solution is $\beta = S_\lambda(y)$, where S_λ is the soft-thresholding operator:

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \le y_i \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}, \quad i = 1,\ldots,n$$

Check: from last slide, subgradient optimality conditions are

$$\begin{cases} y_i - \beta_i = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |y_i - \beta_i| \le \lambda & \text{if } \beta_i = 0 \end{cases}$$

문제 : y와 최대한 가까우면서 너무 크지 않은 β를 찾는것. 직접적인 사용처라기 보다는 보조적으로 사용되는 알고리즘.

: 만약 $\min \frac{1}{2}\|y-\beta\|_2^2$의 경우 $y=\beta$로 두는 것이 최적

$\min \|\beta\|_1$의 경우 $\beta=0$ 일시 최적

따라서 β는 $0 \sim y$ 사이의 어떤 값이 될 것

ex). $y \in \mathbb{R}^1$ 이라고 가정.

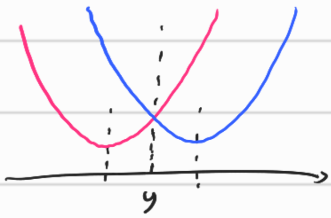
$$f(\beta) = \frac{1}{2}(y-\beta)^2 + \lambda|\beta|$$

$$\approx (y-\beta)^2 + 2\lambda|\beta|$$

$$= \beta^2 - 2y\beta + y^2 + 2\lambda|\beta|$$

$$= \begin{cases} \text{Case 1. } \beta \geq 0 \Rightarrow \beta^2 - 2(y-\lambda)\beta + y^2 \Rightarrow \left(\beta^* = (y-\lambda) \text{ 일때 } y^2-(y-\lambda)^2\right) \\ \text{Case 2. } \beta < 0 \Rightarrow \beta^2 - 2(y+\lambda)\beta + y^2 \Rightarrow \left(\beta^* = (y+\lambda) \text{ 일때 } y^2-(y+\lambda)^2\right) \end{cases}$$

이 두값과 0의 관계에따라서 최소값이 결정.



$\Rightarrow$ 규칙을 따르기에는 too 복잡 (cf. 실습-1. soft thresholding)

$\Rightarrow$ Subgradient Optimality Condition을 쓰자.

$$f(\beta) = \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|\beta\|_1$$

$$\partial f(\beta) = \partial\left(\frac{1}{2}\|y-\beta\|_2^2 + \lambda\|\beta\|_1\right)$$

$$= \partial\left(\frac{1}{2}\|y-\beta\|_2^2\right) + \lambda\,\partial(\|\beta\|_1) \quad \text{(by Additivity and Multiplicity)}$$

$$= -(y-\beta) + \lambda\,\partial(\|\beta\|_1) \quad (\partial f = \nabla f \text{ if } \nabla f \text{ exists})$$

(Let $f(v) = v^T v$, ($\Leftrightarrow \frac{\partial f(v)}{\partial v} = v$), then.

$$\frac{\partial}{\partial\beta}\|y-\beta\|_2^2 = \frac{\partial}{\partial\beta} f(y-\beta) = \frac{\partial f(y-\beta)}{\partial(y-\beta)}\frac{\partial(y-\beta)}{\partial\beta}$$

$$= (y-\beta)\cdot -1$$

$$= -(y-\beta))$$

$$= \begin{cases} -(y-\beta) + \lambda & \text{if } \beta > 0 \quad \text{Case 1.} \\ -(y-\beta) + \lambda \cdot v \ \{v \in [-1,1]\} & \text{if } \beta = 0 \quad \text{Case 2.} \\ -(y-\beta) - \lambda & \text{if } \beta < 0 \quad \text{Case 3.} \end{cases}$$

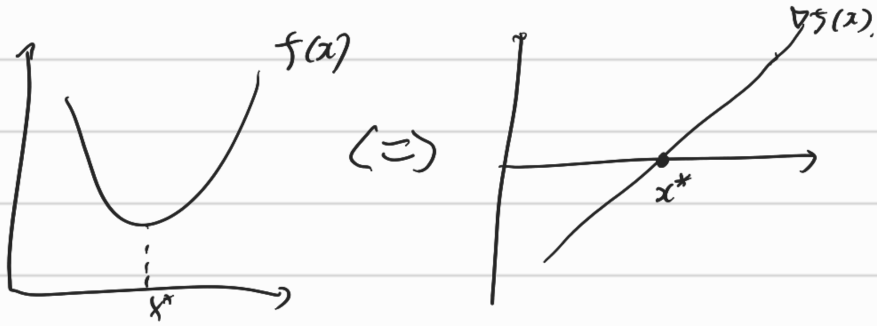
If β^* satisfies $0 \in \partial f(\beta^*)$, β^* is global minima

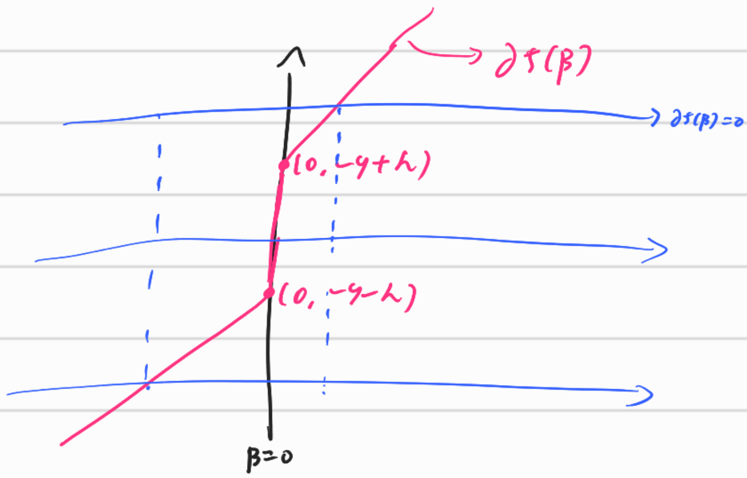
Case 1. $0 = -(y-\beta) + \lambda \Rightarrow \beta = y - \lambda$ if $\beta = y-\lambda > 0$ {i.e. $y > \lambda$}

Case 2. $0 \in -(y-\beta) + \lambda v \Rightarrow \beta \in \lambda v - y = \{[-\lambda-y, \lambda-y]\}$ if $\beta = 0$. {ie $|y| < \lambda$}

Case 3. $0 = -(y-\beta) - \lambda \Rightarrow \beta = y + \lambda$ if $\beta = y+\lambda < 0$ {i.e. $y < -\lambda$}

Remark that to find critical point (변곡점) of $f(x)$, we have to find a point (x; $\nabla f(x) = 0$)
(i.e. a point that $f'(x)$ and x-axis ($y=0$ line) meet)

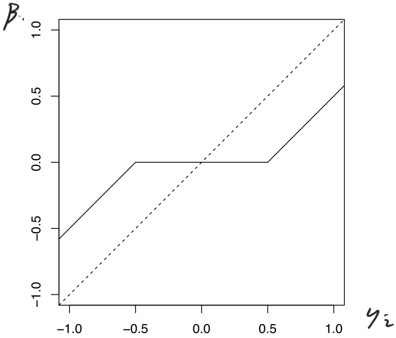




Now plug in $\beta = S_\lambda(y)$ and check these are satisfied:

- When $y_i > \lambda$, $\beta_i = y_i - \lambda > 0$, so $y_i - \beta_i = \lambda = \lambda \cdot 1$
- When $y_i < -\lambda$, argument is similar
- When $|y_i| \le \lambda$, $\beta_i = 0$, and $|y_i - \beta_i| = |y_i| \le \lambda$

Soft-thresholding in one variable:



Example: distance to a convex set

Recall the distance function to a closed, convex set C:

$$\mathrm{dist}(x, C) = \min_{y \in C} \ \|y - x\|_2$$

This is a convex function. What are its subgradients?

Write $\mathrm{dist}(x, C) = \|x - P_C(x)\|_2$, where $P_C(x)$ is the projection of x onto C. It turns out that when $\mathrm{dist}(x, C) > 0$,

$$\partial \mathrm{dist}(x, C) = \left\{ \frac{x - P_C(x)}{\|x - P_C(x)\|_2} \right\}$$

Only has one element, so in fact $\mathrm{dist}(x, C)$ is differentiable and this is its gradient

We will only show one direction, i.e., that

$$\frac{x - P_C(x)}{\|x - P_C(x)\|_2} \in \partial \mathrm{dist}(x, C)$$

Write $u = P_C(x)$. Then by first-order optimality conditions for a projection,

$$(x-u)^T(y-u) \le 0 \quad \text{for all } y \in C$$

Hence

$$C \subseteq H = \{y : (x-u)^T(y-u) \le 0\}$$

Claim:

$$\mathrm{dist}(y, C) \ge \frac{(x-u)^T(y-u)}{\|x-u\|_2} \quad \text{for all } y$$

Check: first, for $y \in H$, the right-hand side is ≤ 0

Now for $y \notin H$, we have $(x-u)^T(y-u) = \|x-u\|_2\|y-u\|_2 \cos\theta$ where θ is the angle between $x-u$ and $y-u$. Thus

$$\frac{(x-u)^T(y-u)}{\|x-u\|_2} = \|y-u\|_2 \cos\theta = \mathrm{dist}(y,H) \le \mathrm{dist}(y,C)$$

as desired

Using the claim, we have for any y

$$\begin{aligned}\mathrm{dist}(y,C) &\ge \frac{(x-u)^T(y-x+x-u)}{\|x-u\|_2} \\ &= \|x-u\|_2 + \left(\frac{x-u}{\|x-u\|_2}\right)^T (y-x)\end{aligned}$$

Hence $g = (x-u)/\|x-u\|_2$ is a subgradient of $\mathrm{dist}(x,C)$ at x

References and further reading

- S. Boyd, Lecture notes for EE 264B, Stanford University, Spring 2010-2011
- R. T. Rockafellar (1970), “Convex analysis”, Chapters 23–25
- L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011-2012