Ch5 - Optimization

2020年11月23日 19:10

Ch. 5. Optimization

We have seen that many data analysis tasks are formulated as an optimization problem

unconstrained optimization

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

or

constrained optimization

$$\min_{x\in\mathbb{R}^n} f(x)$$
$$\text{s.t. } g_i(x)\le 0 \quad i=1,2,\cdots,p$$
$$h_i(x)=0 \quad i=1,2,\cdots,q.$$

This chapter studies the optimality condition for these optimization problems. We consider optimization in $\mathbb{R}^n$ for simplicity. Most of the results in this chapter can be extended to Hilbert spaces H.

Smooth = differentiable

§ 5.1 Smooth Unconstrained Optimization.

Consider unconstrained optimization

$$\min_{x\in\mathbb{R}^n} f(x). \qquad (OPT)$$

We assume $f(x)$ is differentiable.

- Solvability of (OPT).

  We say $x^{(*)}$ is a solution of (OPT) if Def of optimization

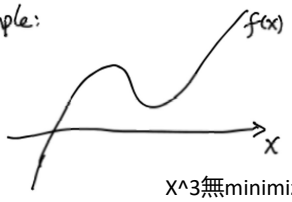
  $$f(x^{(*)}) \le f(x) \qquad \forall\, x\in\mathbb{R}^n.$$

  In this case, we write $x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x)$.

  We also call $x^{(*)}$ a global minimizer of f in $\mathbb{R}^n$.

  - The existence of a solution of (OPT) is NOT guaranteed.

    Example:

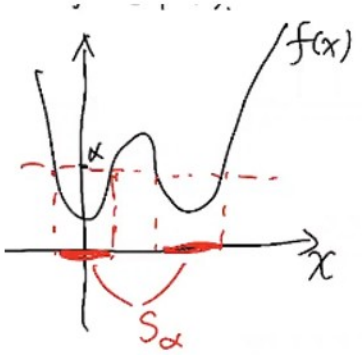
    X^3無minimizer

    $f(x)=x^3, \quad x\in\mathbb{R}$

    $f(x)=\langle a,x\rangle, \quad a,x\in\mathbb{R}^n$

    ⋮

正負x飛去正無限嘅時候，function value都要飛去正無限



if a sequence converge to x, it function value converge to f(x) => continuous

- When there exists a solution of (OPT) ?

Thm: If f is continuous (i.e., $f(x^{(n)}) \to f(x)$ as $x^{(n)} \to x$)(1)
and coecive (i.e., $f(x^{(n)}) \to +\infty$ as $\|x^{(n)}\| \to +\infty$)(2)
Then there exists at least a solution $x^{(*)}$ of (OPT).

proof. Consider the sets $S_\alpha = \{x \mid f(x) \le \alpha\}$ for a given $\alpha \in \mathbb{R}$.

Consider a convergent seq in level set

Use (1) to prove

S_α is closed: Let $x^{(n)} \to x$ and $\{x^{(n)}\} \subset S_\alpha$. Then $f(x^{(n)}) \le \alpha$.
By the continuity, $f(x) \le \alpha \Rightarrow x \in S_\alpha$.

Close set:
For any convergent seq in the set
The limit is also in the set

Use (2) to prove

S_α is bounded: Suppose S_α is unbounded, i.e., $\exists \{x^{(n)}\} \subset S_\alpha$ s.t. $\|x^{(n)}\| \to +\infty$. Then coecivity implies $f(x^{(n)}) \to +\infty$.
This contradicts with $f(x^{(n)}) \le \alpha$.

Bounded: norm of vectors on S_{alpha} is smaller than a constant

Therefore, S_α is bounded and closed for any $\alpha \in \mathbb{R}$.
We choose α s.t. S_α is non-empty By Weierstrass's theorem, any continuous function on a bounded and closed set must have a minimimizer, So, $\min_{x \in S_\alpha} f(x)$ has a solution, which is also a solution of $\min_{x \in \mathbb{R}^n} f(x)$. ⊠

For a continuous function on a closed And bounded and non-empty set, Must have max and min

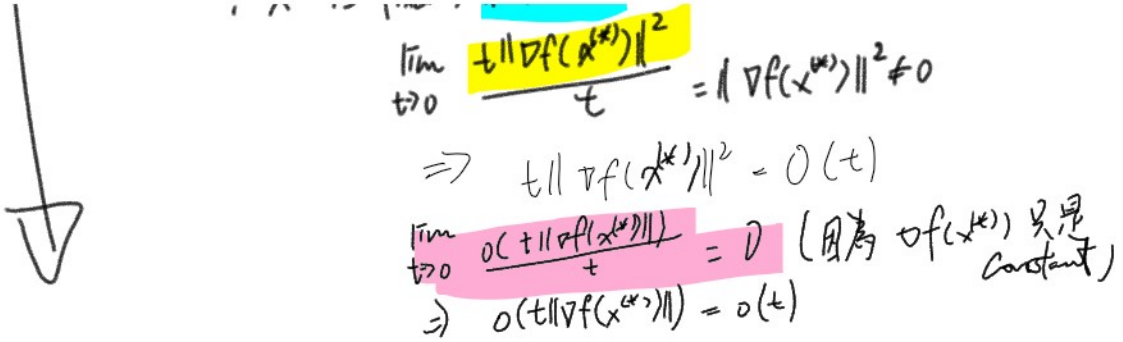
- Necessary condition for optimality.

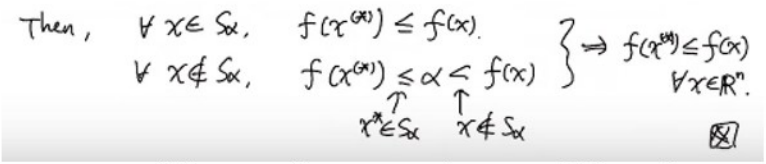
Theorem: Assume f is differentiable at $x^{(*)}$. Then,
$$x^{(*)} = \arg\min_{x \in \mathbb{R}^n} f(x) \Rightarrow \nabla f(x^{(*)}) = 0$$

proof. By expansion, Affine approximation
$$f(x) = f(x^{(*)}) + \langle \nabla f(x^{(*)}), x - x^{(*)} \rangle + o(\|x - x^{(*)}\|_2)$$
Suppose $\nabla f(x^{(*)}) \neq 0$.
Then choose $\tilde{x} = x^{(*)} - t \nabla f(x^{(*)})$ with $t > 0$ gives
$$f(\tilde{x}) = f(x^{(*)}) - t \|\nabla f(x^{(*)})\|^2 + o(|t| \|\nabla f(x^{(*)})\|)$$
Since $\|\nabla f(x^*)\|$ is a constant of $t > 0$, $o(|t| \|\nabla f(x^*)\|) = o(t)$.?

$\because x^{(*)}$ is fixed, $\|\nabla f(x^{(*)})\| \neq 0$ is a constant

$$\lim \frac{t \|\nabla f(x^{(*)})\|^2}{\quad} = \|\nabla f(x^{(*)})\|^2 \neq 0$$





如果佢唔係level set,甘姐係一早都超過alpha

呢到係個proof嘅收尾，prove 呢個f(x(*))一定係全局最細

上述所指只是充分條件，非必要條件

姐係話，唔需要滿足呢兩個都可以有minimizer嘅

例如discontinuous function都可以有minimizer

Reverse of this theorem is not true

上面嘅proof係話三舊野入面，第二舊好近constant, 第三舊好近0，如果揀個足夠細嘅t,就可以chok到第二舊點都大過第三舊

Also, $t\|\nabla f(x^*)\|^2 = O(t)$ (big O of t)

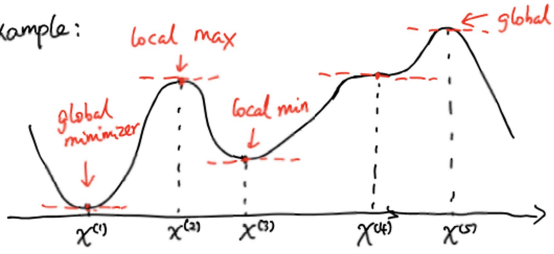
$\Rightarrow$ by choosing a sufficiently small t, $t\|\nabla f(x^{(*)})\|^2 > o\left(|t|\|\nabla f(x^{(*)})\|\right)$,

This implies $f(\tilde{x}) < f(x^{(*)})$, which contradicts with $x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x)$. ⊠

The condition $\nabla f(x^{(*)})=0$ is only a necessary condition.

The reverse "$\nabla f(x^{(*)})=0 \Rightarrow x^{(*)}=\arg\min_{x\in\mathbb{R}^n} f(x)$" is generally NOT true.

Example:

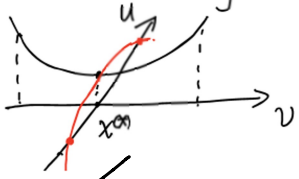


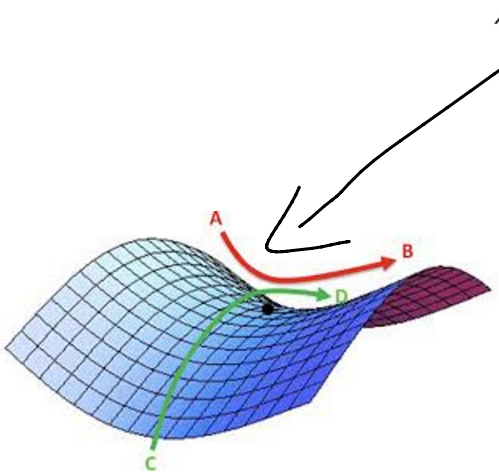
Only $x^{(1)}$ is a global minimizer, though the gradient at $x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}$ are also 0.

From this example, we see that $x^{(*)}$ with $\nabla f(x^{(*)})=0$ can be

- Global minimizer. i.e., $x^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x)$ (see $x^{(1)}$)
- Local minimizer, i.e.,
  $\exists\, \varepsilon$, s.t. $f(x^{(*)}) \le f(x)$ $\forall\, x: \|x-x^{(*)}\|\le\varepsilon$. (see $x^{(2)}$)
- global max, i.e., $\forall\, x\in\mathbb{R}^n$, $f(x^{(*)}) \ge f(x)$, (See $x^{(5)}$)
- Local max, i.e, $\exists\, \varepsilon$. s.t. $f(x^{(*)}) \ge f(x)$ $\forall\, x: \|x-x^{(*)}\|\le\varepsilon$ (See $x^{(3)}$)
- Saddle points (only for $\mathbb{R}^n$ with $n\ge 2$), i.e.,
  $\exists\, u, v \in \mathbb{R}^n$ s.t. $f(x^{(*)}) \ge f(x^{(*)}+tu)$ for all $|t|\le\varepsilon$.
  and $f(x^{(*)}) \le f(x^{(*)}+tv)$ for all $|t|\le\varepsilon$.

(i.e., $f(x^{(*)})$ is a local min along v and a local max along u)





Algebraicly, we checked

linear functions $\subseteq$ affine functions $\subseteq$ convex functions

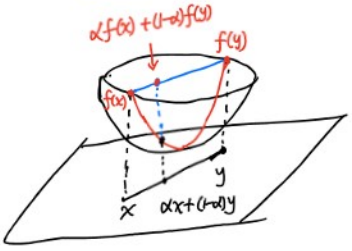
- None of the above (see $x^{(4)}$)

下面就係從f入手，穩d gradient已經滿足足夠以及必要條件作爲minimizer嘅function (i.e. we want to find f s.t. $\nabla f(x^{(4)}) = 0 \iff x^{(4)} = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x)$

- Sufficient condition for optimality

Convexity : A function $f: \mathbb{R}^n \to \mathbb{R}$ is convex if

$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \forall x, y \in \mathbb{R}^n, \ \alpha \in [0,1]$



The secant connecting $(x, f(x))$, $(y, f(y))$ is above the function graph.

idea = make gradient to be monotonically increasing(in 2d only)

Example 1: $f(x) = x^2$, $x \in \mathbb{R}$

$f(\alpha x + (1-\alpha) y)$

$= (\alpha x + (1-\alpha) y)^2$

$= \alpha^2 x^2 + 2\alpha(1-\alpha) xy + (1-\alpha)^2 y^2$

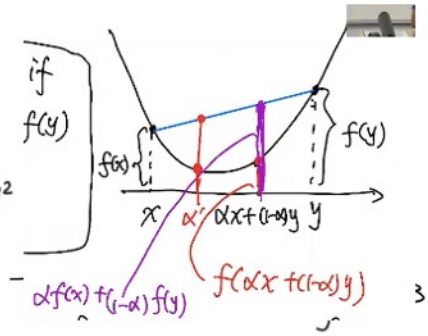
$= \alpha x^2 + (1-\alpha) y^2 + (\alpha^2 - \alpha) x^2$

$\quad + (\alpha^2 - \alpha) y^2 + 2\alpha(1-\alpha) xy$

$= \alpha x^2 + (1-\alpha) y^2 - \alpha(1-\alpha)(x^2 + y^2 - 2xy)$

$= \alpha x^2 + (1-\alpha) y^2 - \alpha(1-\alpha)(x-y)^2$

$\leq \alpha x^2 + (1-\alpha) y^2 = \alpha f(x) + (1-\alpha) f(y)$



Example 2: $f(x) = \|x\|^2$

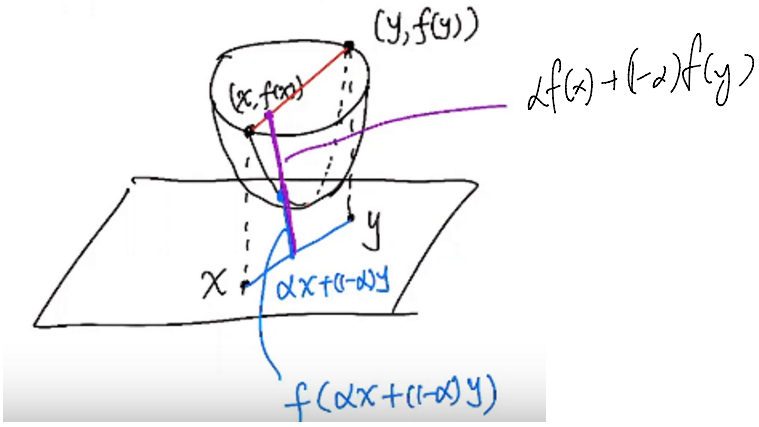
$f(\alpha x + (1-\alpha) y) = \|\alpha x + (1-\alpha) y\|^2 = \alpha^2 \|x\|^2 + (1-\alpha)^2 \|y\|^2 + 2\alpha(1-\alpha) \langle x, y \rangle$

$= \alpha \|x\|^2 + (1-\alpha) \|y\|^2 + 2\alpha(1-\alpha) \langle x, y \rangle + (\alpha^2 - \alpha) \|x\|^2 + (\alpha^2 - \alpha) \|y\|^2$

$= \alpha \|x\|^2 + (1-\alpha) \|y\|^2 - \alpha(1-\alpha) (\|x\|^2 + \|y\|^2 - 2\langle x, y \rangle)$

$= \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha) \|x-y\|^2 \leq \alpha f(x) + (1-\alpha) f(y)$

Example 3: $f(x) = \|x\|$, where $\|x\|$ is a norm on $\mathbb{R}^n$ (e.g. p-norm, 1-norm, ...)



$$f(\alpha x+(1-\alpha)y) = \|\alpha x+(1-\alpha)y\| \le \|\alpha x\|+\|(1-\alpha)y\|$$
$$\le \alpha\|x\|+(1-\alpha)\|y\| \qquad \forall\ \alpha\in[0,1],\ x,y\in\mathbb{R}^n$$

Therefore, any norm function is convex.

Example 4: Any affine function f is convex, since

$$f(\alpha x+(1-\alpha)y) = \alpha f(x)+(1-\alpha)f(y) \le \alpha f(x)+(1-\alpha)f(y).$$

$\forall\ \alpha\in[0,1]$, $x,y\in\mathbb{R}^n$.

Convex function = extension of affine function

Example 5: Let $f_1,\cdots,f_n$ are convex, then $f=\sum_{i=1}^n c_i f_i$, $c_i\ge 0$, is convex.

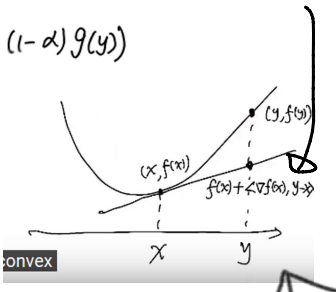
$$f(\alpha x+(1-\alpha)y)=\sum_{i=1}^n c_i f_i(\alpha x+(1-\alpha)y) \le \sum_{i=1}^n c_i\left(\alpha f_i(x)+(1-\alpha)f_i(y)\right)$$
$$= \alpha\sum_{i=1}^n c_i f_i(x)+(1-\alpha)\sum_{i=1}^n c_i f_i(y) = \alpha f(x)+(1-\alpha)f(y)$$

Example 6: Let f be convex and g be affine.

Then $f\circ g$ is convex.

$$(f\circ g)(\alpha x+(1-\alpha)y) = f\left(g(\alpha x+(1-\alpha)y)\right) = f\left(\alpha g(x)+(1-\alpha)g(y)\right)$$
$$\le \alpha f(g(x))+(1-\alpha)f(g(y)).$$

Affine approximation



Theorem: If f is convex and differentiable, then

$$x^{(*)}=\arg\min_{x\in\mathbb{R}^n} f(x) \iff \nabla f(x^{(*)})=0.$$

Extra example

To prove the theorem, we need a Lemma, which is also useful later.

Lemma: If f is differentiable, then

$$f \text{ is convex} \iff f(y)\ge f(x)+\langle\nabla f(x), y-x\rangle \qquad \forall\ x,y.$$

proof. We prove the one variable case, i.e., $f:\mathbb{R}\to\mathbb{R}$.

"$\Rightarrow$". By convexity, $f(\alpha x+(1-\alpha)y)=f(x+(1-\alpha)(y-x))\le \alpha f(x)+(1-\alpha)f(y)$

This implies $f(y)\ge f(x)+\dfrac{f(x+(1-\alpha)(y-x))-f(x)}{(1-\alpha)(y-x)}(y-x)$

Let $\alpha\to 1$, then $f(y)\ge f(x)+f'(x)\cdot(y-x)$

"$\Leftarrow$" Choose $x\ne y$ and $\alpha\in[0,1]$.

Consider $z=\alpha x+(1-\alpha)y$

both are non-negative

Consider $z=\alpha x+(1-\alpha)y$.

"⇐" Choose $x \neq y$ and $\alpha \in [0,1]$.

Consider $z = \alpha x + (1-\alpha) y$

Then $f(x) \geq f(z) + f'(z)(x-z)$ (1) $\times \alpha$

$f(y) \geq f(z) + f'(z)(y-z)$ (2) $\times (1-\alpha)$ +

$\alpha f(x) + (1-\alpha) f(y) \geq f(z) + f'(z)\left(\alpha x + (1-\alpha) y - z\right)$

0

So, $f(\alpha x + (1-\alpha) y) = f(z) \leq \alpha f(x) + (1-\alpha) f(y)$

$\Rightarrow f$ is convex.

both are nonnegative

Consider $z = \alpha x + (1-\alpha) y$.

$x - z$ (1)

姐係話，f is convex if and only if it is higher than its tangent plane everywhere

Example 7: Let $f: \mathbb{R}^n \to \mathbb{R}$ be convex.

Then $g(x) = f(Ax+b)$, where A is a matrix, b is a vector

is convex.

proof. $g(\alpha x + (1-\alpha) y) = f\left(A(\alpha x + (1-\alpha) y) + b\right)$

$= f\left(\alpha Ax + (1-\alpha) Ay + \alpha b + (1-\alpha) b\right)$

$= f\left(\alpha (Ax+b) + (1-\alpha)(Ay+b)\right)$

$\leq \alpha f(Ax+b) + (1-\alpha) f(Ay+b)$

Affine transformation

"⇒" Prof: By convexity,

$$f(\alpha x + (1-\alpha) y) = f\left(x + (1-\alpha)(y-x)\right) \leq \alpha f(x) + (1-\alpha) f(y)$$

$$\therefore \underbrace{(1-\alpha) f(x) - f(x)}_{-\alpha f(x)} + \underbrace{f\left(x + (1-\alpha)(y-x)\right)}_{\text{LHS}} \leq \underbrace{(1-\alpha) f(y)}_{\text{RHS}}$$

choose $\alpha \in (0,1)$,

$$f(y) \geq f(x) + \frac{f\left(x + (1-\alpha)(y-x)\right) - f(x)}{1-\alpha}$$

(兩邊同時除1-alpha)

let $\alpha \to 1$, we obtain

$$f(y) \geq f(x) + \frac{d}{dt} f\left(x + t(y-x)\right)\Big|_{t=0}$$

$$= f(x) + f'(x)(y-x)$$

~~$f(y) \geq f(z) + f'(z)(y-z)$ (2)~~

~~(1) $\times \alpha$ + (2) $\times (1-\alpha)$ $\Rightarrow f(z) \leq \alpha f(x) + (1-\alpha) f(y)$.~~

Next, we prove the multivariable case, i.e., $f: \mathbb{R}^n \to \mathbb{R}$.

(1) $\times\alpha$ + (2) $\times(1-\alpha)$ $\Rightarrow$ $f(z) \le \alpha f(x) + (1-\alpha) f(y)$.

Next, we prove the multivariable case, i.e., $f: \mathbb{R}^n \to \mathbb{R}$.

"$\Rightarrow$". Consider a function $g(t) = f(tx + (1-t)y)$, $t \in \mathbb{R}$.

So $g'(t) = \langle \nabla f(tx+(1-t)y), x-y \rangle$ by chain rule.

Since f is convex,

$$g(\alpha s + (1-\alpha)t) = f\big((\alpha s + (1-\alpha)t)x + (1-\alpha s - (1-\alpha)t)y\big)$$
$$= f\big(\alpha(sx + (1-s)y) + (1-\alpha)(tx + (1-t)y)\big)$$
$$\le \alpha f(sx + (1-s)y) + (1-\alpha) f(tx + (1-t)y) = \alpha g(s) + (1-\alpha) g(t)$$

which implies $g(t)$ is convex.

The result from one variable case leads to

$$g(0) \ge g(1) + g'(1) \cdot (-1),$$

i.e., $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle$

"$\Leftarrow$". Choose $x \ne y$, $\alpha \in [0,1]$, consider $z = \alpha x + (1-\alpha) y$.

Then $f(x) \ge f(z) + \langle \nabla f(z), x - z \rangle$ (1)

$f(y) \ge f(z) + \langle \nabla f(z), y - z \rangle$ (2)

(1) $\times\alpha$ + (2) $\times(1-\alpha)$ $\Rightarrow$ $f(z) \le \alpha f(x) + (1-\alpha) f(y)$. ⊠

$\alpha f(x) + (1-\alpha) f(y) \ge f(z) + \langle \nabla f(z), (\alpha x + (1-\alpha) y - z) \rangle$

∵ $f(tx+(1-t)y) = f(y + t(x-y))$,

by chain rule,

$g'(t) = \langle \nabla f(tx+(1-t)y, \frac{d}{dt}(y + t(x-y)) \rangle$

$g'(t) = \langle \nabla f(tx+(1-t)y, x-y \rangle$

1D result:

$g(s) \ge g(t) + g'(t)\cdot(s-t)$

$\forall s, t \in \mathbb{R}$

$\Rightarrow$ choose $s=1, t=0$,

$g(0) \ge g(1) + g'(1)(-1)$

$f(y) \ge f(x) + \langle \nabla f(x), y-x \rangle$

With the lemma, now we can prove the theorem.

proof of the theorem: We prove only $\nabla f(x^{(*)}) = 0 \Rightarrow x^{(*)} = \arg\min_x f(x)$

Since f is convex and differentiable, for any $x \in \mathbb{R}^n$,

$$f(x) \ge f(x^{(*)}) + \langle \nabla f(x^{(*)}), x - x^{(*)} \rangle$$

By assumption, $\nabla f(x^{(*)}) = 0$

$\Rightarrow$ $f(x) \ge f(x^{(*)})$, i.e., $x^{(*)} = \arg\min f(x)$. ⊠

呢到用翻啱啱嘅lemma

- When the global minimizer is unique?

Strictly convex function: A function $f: \mathbb{R}^n \to \mathbb{R}$ is strictly convex if

$$f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y) \quad \forall\, x \ne y,\ \alpha \in (0,1)$$

- Example 1: $f(x) = x^2$, $x \in \mathbb{R}$.

$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha)(x-y)^2 \quad \forall\, x, y \in \mathbb{R}^n,\ \alpha \in [0,1]$$

Therefore, $f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y)$ if $x \ne y$, $\alpha \ne 0, \alpha \ne 1$, $\alpha \in (0,1)$

Thus f is strictly convex.

所以, $f(x) = |x|$ 都不是 strictly convex,

strictly convex,
choose $x=1, y=2, \alpha \in (0,1)$

$f(\alpha x+(1-\alpha)y) = |\alpha x+(1-\alpha)y| = \alpha|x|+(1-\alpha)|y|$
$= \alpha f(x)+(1-\alpha)f(y)$
$\Rightarrow$ NOT strictly convex.

如果function唔係flat
就不是strictly convex

Therefore, $f(\alpha x+(1-\alpha)y) < \alpha f(x)+(1-\alpha)f(y)$ if $x \neq y$, $\alpha \neq 0, \alpha \neq 1$, $\alpha \in (0,1)$

Thus, f is strictly convex.

- Example 2: $f(x) = \|x\|_2^2$, $x \in \mathbb{R}^n$, is strictly convex.

Example 4: $f(x) = \|x\|_1$, $x \in \mathbb{R}^n$ is convex but NOT strictly convex

— Because $\|x\|_1$ is a norm on $\mathbb{R}^n$, $f(x) = \|x\|_1$ is convex.

— Let $x = e_1$, $y = e_2$, and $\alpha \in (0,1)$

$f(\alpha x+(1-\alpha)y) = \|\alpha x+(1-\alpha)y\|_1 = \left\| \begin{bmatrix} \alpha \\ 1-\alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\|_1$

$= |\alpha| + |1-\alpha| = 1$

$\alpha f(x)+(1-\alpha)f(y) = \alpha \|x\|_1 + (1-\alpha)\|y\|_1 = \alpha + (1-\alpha) = 1$

$\Rightarrow f(\alpha x+(1-\alpha)y) = \alpha f(x)+(1-\alpha)f(y)$

- Example 4: If $f_1, f_2, \cdots, f_n$ are strictly convex, then $f = \sum_{i=1}^{n} c_i f_i$, where $c_i \geq 0$ for all i and not all c_i's are 0, is strictly convex.

Theorem: Assume f is strictly convex. Then the solution of $\min_{x \in \mathbb{R}^n} f(x)$ is unique if it exists.

proof. Suppose there are at least two solutions $x^{(*)}$, $y^{(*)}$. Then $f(x^{(*)}) = f(y^{(*)})$

Consider $z = \alpha x^{(*)} + (1-\alpha) y^{(*)}$ with $\alpha \in (0,1)$.

Then $z \neq x^{(*)}$, $z \neq y^{(*)}$. Moreover,

$f(z) = f(\alpha x^{(*)} + (1-\alpha) y^{(*)})$

$< \alpha f(x^{(*)}) + (1-\alpha) f(y^{(*)}) = \alpha f(x^{(*)}) + (1-\alpha) f(x^{(*)})$

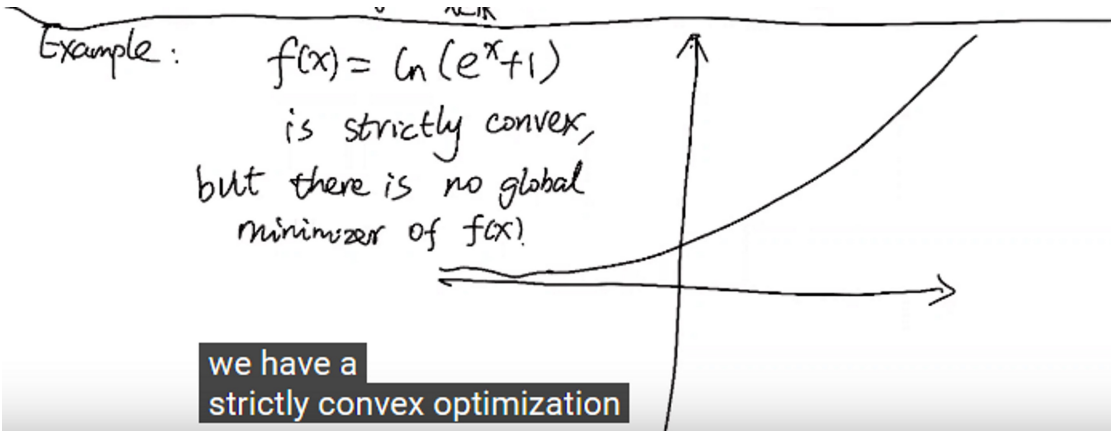
$= f(x^{(*)})$,

Strictly convex

Contradict jor, f(z) should not smaller than f(x(*))

i.e., $f(z) < f(x^{(*)})$, which contradicts with $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} f(x)$.

Remark: Even if f is convex/strictly convex, the existence of a solution of $\min_{x \in \mathbb{R}^n} f(x)$ is NOT guaranteed.



i.e., $f(z) < f(x^{(*)})$, which contradicts with $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} f(x)$. ⊗

- Gradient Descent

The simplest algorithm for finding a solution of $\min_{x \in \mathbb{R}^n} f(x)$ is gradient descent in case that f is differentiable.

Let $x^{(k)}$ be the current estimation.

Let us find $x^{(k+1)}$ such that $f(x^{(k+1)}) \leq f(x^{(k)})$

Given $x^{(k)}$, we approximate $f(x)$ linearly via Taylor's expansion.

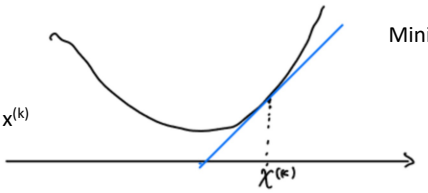
$$f(x) \approx f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle$$

但係用這個minimizer有兩個問題:

1.linear approximation無minimizer

2. 如果x與$x^{(k)}$距離好遠，個估計會唔準確

所以我地minimize f in a small neighborhood of $x^{(k)}$



Minimize an approximation of f

This approximation is accurate only when x is close to $x^{(k)}$.

Therefore, instead of $\min_{x \in \mathbb{R}^n} f(x)$, we minimize the linear approximation and the distance of x to $x^{(k)}$ simultaneously

Standard gradient descent

To achive this, we solve

$$x^{(k+1)} = \arg\min_{x \in \mathbb{R}^n} f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\alpha_k} \|x - x^{(k)}\|_2^2$$

呢到都可以用1 norm

to make the linear approximation small

to make x close to $x^{(k)}$

where $\alpha_k > 0$ is a parameter to balance the two terms.

Penalize term

Let $F(x) = f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\alpha_k} \|x - x^{(k)}\|_2^2$

It is easily checked that F is convex. Thus,

$$\min_{x \in \mathbb{R}^n} F(x) \iff \nabla F(x) = 0$$

$$\iff \nabla f(x^{(k)}) + \frac{1}{\alpha_k}(x - x^{(k)}) = 0$$

$$\iff x = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

如果alpha_k好細，甘就係倍大呢個penalize term

F(x) is continuous, 因爲inner product and 2 norm都係continuous, it is coercive

coercive :

① F is continous, and coercive.

— Continuity is obvious

— Coercivity:

$$F(x) = f(x^{(k)}) + \frac{1}{\alpha_k}\left(\frac{1}{2}\|x-x^{(k)}\|_2^2 + \langle x-x^{(k)}, \alpha_k \nabla f(x^{(k)})\rangle\right)$$

試下爆呢行，會等於上面個行

$$= f(x^{(k)}) + \frac{1}{\alpha_k}\left(\frac{1}{2}\|x-x^{(k)} + \alpha_k \nabla f(x^{(k)})\|_2^2 - \frac{1}{2}\alpha_k^2 \|\nabla f(x^{(k)})\|_2^2\right)$$

$$= \frac{1}{2\alpha_k}\|x - (x^{(k)} - \alpha_k \nabla f(x^{(k)}))\|_2^2 - \frac{\alpha_k}{2}\|\nabla f(x^{(k)})\|_2^2 + f(x^{(k)})$$

拆開

$$\geq \frac{1}{2\alpha_k}\left(\|x\|_2 - \|x^{(k)} - \alpha_k \nabla f(x^{(k)})\|_2\right)^2 - \frac{\alpha_k}{2}\|\nabla f(x^{(k)})\|_2^2 + f(x^{(k)})$$

Obviously, when $\|x\|_2 \to +\infty$, $F(x) \to +\infty$.

So, $\min_{x\in\mathbb{R}^n} F(x)$ has at least a solution.

② F is strictly convex,

because

$-f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)}\rangle$ is convex in x.

$$-\frac{1}{2\alpha_k}\|x-x^{(k)}\|_2^2 = \frac{1}{2\alpha_k}\|x\|_2^2 - \frac{1}{\alpha_k}\langle x, x^{(k)}\rangle + \frac{1}{2\alpha_k}\|x^{(k)}\|_2^2$$

$\frac{1}{2\alpha_k}\|x\|_2^2$: strictly convex; $-\frac{1}{\alpha_k}\langle x, x^{(k)}\rangle$: convex (∵ affine)

is strictly convex in x.

So F = convex function + strictly convex function

is strictly convex.

Thus, $\min_{x\in\mathbb{R}^n} F(x)$ has a unique solution

$x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} F(x)$ is well-defined

3. Since F is convex and differentiable

$$x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} F(x) \iff \nabla F(x^{(k+1)}) = 0$$

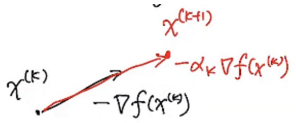
Since

$$\nabla F(x) = \nabla f(x^{(k)}) + \frac{1}{\alpha_k}(x - x^{(k)})$$

So, $\nabla F(x^{(k+1)}) = 0 \iff \nabla f(x^{(k)}) + \frac{1}{\alpha_k}(x^{(k+1)} - x^{(k)}) = 0$

$$\iff \boxed{x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})}$$

Update $x^{(k)}$ to $x^{(k+1)}$



We obtain an algorithm:

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}), \qquad k=0,1,2,\ldots$$

which is known as Gradient Descent algorithm.

— The parameter $\alpha_k > 0$ is called a step size in optimization or a learning rate in machine learning.

— How to choose α_k ?

① Exact linear search: Choose

$$\alpha_k = \arg\min_{\alpha \geq 0} f\left(x^{(k)} - \alpha \nabla f(x^{(k)})\right).$$

② Linear search by back-tracking:

- Try a very large α_k.
- Test whether or not α_k is good enough:
  - — Yes. Go to the next iteration.
  - ← No. Decrease α_k and try again.

Example: Armijo-Goldstein backtracking

Choose a large α_0. Choose $\beta < 1$ (e.g. $\beta = 0.9$).

At step k: Discount rate

Use beta to discount

$\alpha_k = \alpha_{k-1}$

$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$

if $f(x^{(k)}) - f(x^{(k+1)}) < \frac{\alpha_k}{2} \|\nabla f(x^{(k)})\|_2^2$

then $\alpha_k = \beta \alpha_k$ and goto

otherwise go to the next iteration

There are many other criteria to test a good step size

Search "Armijo-Goldstein Condition" in google

③ Use a fixed step size. 因爲 $\nabla f(x) = 0$ 就到 minimize 3.

所以可以用gradient黎check下是否acceptable

- Choice of step size α_k. (i.e., $\alpha_k = \alpha$, $\forall k$) $(\alpha > 0)$

如果係convex, 個graph點都大過tangent plane嘅

For simplicity, we assume f is convex.

By convexity, the linear approximation gives a lower bound of $f(x^{(k+1)})$

$$f(x^{(k+1)}) \geq f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle.$$

呢個只係lower bound,但係我地想要嘅係一個upper bound令到$f(x^{(k+1)})$足夠小

To assure $f(x^{(k+1)}) \leq f(x^{(k)})$, we need, however, an upper bound

如果係convex, 個graph點都大過tangent plane嘅

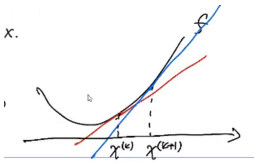
For simplicity, we assume f is convex.

By convexity, the linear approximation gives a lower bound of $f(x^{(k+1)})$

$$f(x^{(k+1)}) \geq f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle.$$

呢個只係lower bound,但係我地想要嘅係一個upper bound令到f(x^(k+1))足夠小

To assure $f(x^{(k+1)}) \leq f(x^{(k)})$, we need, however, an upper bound of $f(x^{(k+1)})$. To this end, we consider linear approximation at $x^{(k+1)}$ as



$$f(x^{(k)}) \geq f(x^{(k+1)}) + \langle \nabla f(x^{(k+1)}), x^{(k)} - x^{(k+1)} \rangle$$

i.e.,
$$\begin{aligned} f(x^{(k+1)}) &\leq f(x^{(k)}) - \langle \nabla f(x^{(k+1)}), x^{(k)} - x^{(k+1)} \rangle \\ &= f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle \\ &\quad + \langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle \end{aligned}$$

$\Leftarrow = f(x^{(k)}) + \langle \nabla f(x^{(k+1)}), x^{(k+1)} - x^{(k)} \rangle$ 加減一步.

相當于考慮另一條tangent line

Since, by gradient descent algorithm,

$$x^{(k+1)} - x^{(k)} = -\alpha \nabla f(x^{(k)}), \quad \forall k.$$

So,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \alpha \|\nabla f(x^{(k)})\|_2^2 + \langle \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$$

To obtain an upper bound of $f(x^{(k+1)})$, we assume

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq M \|x - y\|_2^2 \quad \forall\, x, y \in \mathbb{R}^n$$

(The gradient $\nabla f(x)$ doesn't change too much w.r.t x)

Then,

$$\begin{aligned} f(x^{(k+1)}) &\leq f(x^{(k)}) - \alpha \|\nabla f(x^{(k)})\|_2^2 + M \|x^{(k+1)} - x^{(k)}\|_2^2 \\ &= f(x^{(k)}) - \alpha \|\nabla f(x^{(k)})\|_2^2 + M\alpha^2 \|\nabla f(x^{(k)})\|_2^2 \\ &= f(x^{(k)}) - \alpha(1 - M\alpha) \|\nabla f(x^{(k)})\|_2^2. \end{aligned}$$

To make $f(x^{(k+1)}) < f(x^{(k)})$ for $\nabla f(x^{(k)}) \neq 0$, we require

$$\alpha > 0 \quad \text{and} \quad (1 - M\alpha) > 0 \quad (\text{i.e. } \alpha < \tfrac{1}{M}).$$

Thus, we choose $\alpha \in (0, \frac{1}{M})$.

Indeed, this estimation can be extended to non-convex f and

呢到係講緊，如果grad(f)有時候都會上升嘅話，當然唔可以上升太多啦，
只要relax alpha到2/M就可以了（extenstion, 無講proof）

can be improved to $\alpha \in (0, \frac{2}{M})$.

Furthermore, we have

Theorem: Assume ① f is differentiable and $\exists$ a solution of $\min_{x \in \mathbb{R}^n} f(x)$.

② $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq M \|x - y\|_2^2$, $\forall x, y \in \mathbb{R}^n$.

Theorem: Assume ① f is differentiable and $\exists$ a solution of $\min_{x\in\mathbb{R}^n} f(x)$.

② $\langle \nabla f(x) - \nabla f(y), x-y\rangle \le M\|x-y\|_2^2, \quad \forall x, y \in \mathbb{R}^n$.

③ $\alpha \in (0, \frac{2}{M})$ Reasonable step size

Then, the sequence $\{x^{(k)}\}$ generated by

$$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)}) \quad \text{and } x^{(0)} \in \mathbb{R}^n$$

Satisfies: i). $f(x^{(k+1)}) < f(x^{(k)})$ (the function value decreases)

ii) $\lim_{k\to+\infty} \|\nabla f(x^{(k)})\|_2 = 0$ (the limit has a vanishing gradient)

objective function will be monotonically decreasing

- Since a vanishing gradient is not a sufficient condition for a global minimizer, the gradient descent is NOT guaranteed to find a solution of $\min_{x\in\mathbb{R}^n} f(x)$. It finds only a vanishing gradient with a decreasing function value of f.
- In the special case "f is convex", the gradient descent will finds a solution of $\min_{x\in\mathbb{R}^n} f(x)$, as the global minimizer is equivalent to a vanishing gradient.

If f is non-convex, the gradient descent only gives us a critical point of f.

§5.2 Case Studies of Gradient Descent

§5.2.1. Least Squares

- Recall linear regression:

  Given $(x_1, y_1), \cdots, (x_N, y_N)$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$.

  we want to find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ s.t.

  $$\langle x_i, a \rangle + b \approx y_i, \quad i = 1, \cdots, N.$$

  via minimizing the squares error

  $$\min_{\substack{a \in \mathbb{R}^n \\ b \in \mathbb{R}}} \frac{1}{2} \sum_{i=1}^{N} (\langle x_i, a \rangle + b - y_i)^2$$

  Let $X = \begin{bmatrix} x_1^T & 1 \\ \vdots & \vdots \\ x_N^T & 1 \end{bmatrix} \in \mathbb{R}^{N \times (n+1)}$ $\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$ $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$.

  Then we need to solve the least squares (LS) problem

  $$\min_{\beta \in \mathbb{R}^{n+1}} \frac{1}{2} \|X\beta - y\|_2^2$$

- We consider the standard LS problem

  $$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2, \quad \text{where } A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m.$$

  Let $f(x) = \frac{1}{2} \|Ax - b\|_2^2$.

- $f(x)$ is convex, because:

  Let $f_1(x) = Ax - b$, $f_2(y) = \frac{1}{2}\|y\|_2^2$. Then $f = f_2 \circ f_1$

  Since f_1 is affine and f_2 is convex, f is convex.

- Obviously, f is differentiable. We obtain $\nabla f(x)$ by the following:

  We approximate $f(x)$ by: At any $x \in \mathbb{R}^n$, for any $y \in \mathbb{R}^n$

  $$\begin{aligned} f(y) &= \frac{1}{2}\|Ay - b\|_2^2 = \frac{1}{2}\|(Ax - b) + A(y - x)\|_2^2 \\ &= \frac{1}{2}\|Ax - b\|_2^2 + \langle Ax - b, A(y - x) \rangle + \frac{1}{2}\|A(y - x)\|_2^2 \\ &= f(x) + \langle A^T(Ax - b), y - x \rangle + \frac{1}{2}\|A(y - x)\|_2^2 \end{aligned}$$

  Affine approximation; error

  where we have used the equality $\langle Ax - b, A(y - x) \rangle = \langle A^T(Ax - b), y - x \rangle$.

From linear algebra, $\langle u, Bv \rangle = u^T B v = (u^T B v)^T = v^T B^T u$ ($u^T B v$ is a number)

$= \langle v, B^T u \rangle$

$= \langle B^T u, v \rangle$

u bb equals to b transpose

Generally, we have

$$\langle u, Mv \rangle = \sum_i \sum_j m_{ij} u_i v_j = \sum_j \left(\sum_i m_{ij} u_i \right) v_j = \langle M^T u, v \rangle$$

($\sum_i m_{ij} u_i = (M^T u)_j$)

for any vectors u, v and any matrix M.

Let us estimate the error: (unit l_2)

for any vectors u, v and any matrix M.

Let us estimate the error:

$$\frac{1}{2}\|A(y-x)\|_2^2 = \frac{1}{2}\left\|A\frac{(y-x)}{\|y-x\|_2}\right\|_2^2 \cdot \|y-x\|_2^2$$

$$\leq \frac{1}{2}\left(\max_{\|z\|_2=1}\|Az\|_2^2\right)\|y-x\|_2^2 = \frac{1}{2}\left(\max_{\|z\|_2=1}\|Az\|_2\right)^2\|y-x\|_2^2$$

The quantity $\max_{\|z\|_2=1}\|Az\|_2$ is finite, because

- $g(z)=\|Az\|_2=\sqrt{\sum_i\left(\sum_j a_{ij}z_j\right)^2}$ is continuous
- the set $\{z \mid \|z\|_2=1\}$ is bounded and closed, and non-empty

By Weierstrauss' theorem, $\max_{\|z\|_2=1}\|Az\|_2$ exists and is finite.

Actually $\max_{\|z\|_2=1}\|Az\|_2$ is a norm of $A\in\mathbb{R}^{m\times n}$, denoted by $\|A\|_2$.

Define as M

$M = \max_{\|z\|_2\leq 1}\|Az\|_2 =$

$\left(\max_{\|z\|_2\leq 1}\|Az\|_2^2\right)^2 =$

$\|Az\|_{2\to 2}^2 = \|A\|_2^2$

Bounded closed non-empty set must have max and min

Thus, $\frac{1}{2}\|A(x-y)\|_2^2 \leq \frac{1}{2}\|A\|_2^2\|y-x\|_2^2$, such that

$$0 \leq \lim_{\|y-x\|_2\to 0}\frac{\frac{1}{2}\|A(x-y)\|_2^2}{\|y-x\|_2} \leq \lim_{\|y-x\|_2\to 0}\frac{\frac{1}{2}\|A\|_2^2\|y-x\|_2^2}{\|y-x\|_2} = 0$$

Hence, $\boxed{\nabla f(x) = A^T(Ax-b)}$

- Therefore,

$$x^{(*)} = \arg\min_{x\in\mathbb{R}^n}\frac{1}{2}\|Ax-b\|_2^2 \iff \boxed{A^TAx^{(*)} = A^Tb}$$

called the Normal equation of Least Squares.

- Geometric explanation:

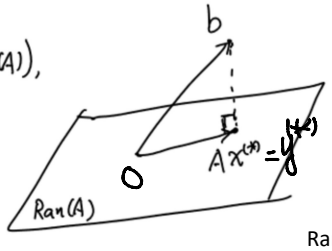
Ax is always in the range of A ($\text{Ran}(A)$),

b is NOT necessarily in $\text{Ran}(A)$

Therefore,

$$\min_{x\in\mathbb{R}^n}\frac{1}{2}\|Ax-b\|_2^2 \iff \min_{y\in\text{Ran}(A)}\frac{1}{2}\|y-b\|_2^2$$

i.e., $Ax^{(*)}$ is the projection of b onto $\text{Ran}(A)$, So,



Ran(A) = subspace in R^m

Assume m > n

Ax in Ran(A)

$$b - Ax^{(*)} \perp \text{Ran}(A),$$

which is the same as

$$\langle Ax^{(*)}-b, Ay\rangle = 0 \quad \forall y\in\mathbb{R}^n.$$

$$\iff \langle A^T(Ax^{(*)}-b), y\rangle = 0 \quad \forall y\in\mathbb{R}^n$$

$$\iff A^T(Ax^{(*)}-b) = 0. \quad A^TAx^{(*)} = A^Tb$$

- It can be shown that: if A^TA is invertible, then $f(x)$ is strictly convex, and therefore the solution of least squares is unique.
- To find $x^{(*)}$, we may use a gradient descent

$$x^{(k+1)} = x^{(k)} - \alpha_k\underbrace{A^T(Ax^{(k)}-b)}_{\nabla f(x)}$$

不能用LU.

迭代 iterative

$b-Ax^{(*)}$ 像 residual

• To find $x^{(k)}$, we may use a gradient descent

$$x^{(k+1)} = x^{(k)} - \alpha_k \underbrace{A^T(Ax^{(k)} - b)}_{\nabla f(x)}$$

residual

To choose a good α_k, we may use "line search", i.e., we set $\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f\left(x^{(k)} - \alpha A^T(Ax^{(k)} - b)\right)$.

In other words, α_k is the optimal step size

Let $g(\alpha) = f(x^{(k)} - \alpha A^T(Ax^{(k)} - b))$.

It can be checked $g(\alpha)$ is convex, and therefore

$$g'(\alpha_k) = 0,$$

which gives $\alpha_k = \dfrac{\|A^T(Ax^{(k)} - b)\|_2^2}{\|AA^T(Ax^{(k)} - b)\|_2^2}$.

This leads to the steepest descent algorithm for least squares

Initialize $x^{(0)}$

for $k = 0, 1, 2, \cdots$

$$g^{(k)} = A^T(Ax^{(k)} - b)$$

$$\alpha_k = \frac{\|g^{(k)}\|_2^2}{\|Ag^{(k)}\|_2^2}$$

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

end

Proj of convex func to line still convex

Also other algo: e.g. conjugate gradient(CG) (but prof said skip it)

- When A^TA is non-invertible, the least squares doesn't have a unique solution. Other restriction on x

  In this case, we usually use regularization techniques.

- Ridge Regression. (use $\|x\|_2^2$ as regularization)

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \frac{\lambda}{2}\|x\|_2^2,$$

where $\lambda>0$ is a regularization parameter.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \frac{\lambda}{2}\|x\|_2^2$

- Since $\frac{\lambda}{2}\|x\|_2^2$ is strictly convex and $\frac{1}{2}\|Ax-b\|_2^2$ is convex, $f(x)$ is strictly convex.

  — $\nabla f(x) = A^T(Ax-b)+\lambda x$

  — There is a unique solution of Ridge Regression, which is given by the solution of

$$\nabla f(x)=0 \iff A^T(Ax-b)+\lambda x=0$$
$$\iff (A^TA+\lambda I)x = A^Tb.$$

- We can solve it by Gradient Descent with exact line search.

- LASSO regression (use $\lambda\|x\|_1$ as regularization)

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1,$$

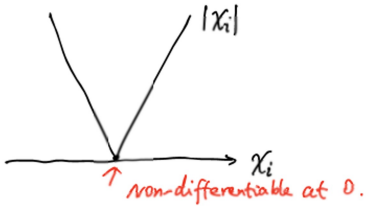
where $\lambda>0$ is a regularization parameter.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$

— $f(x)$ is convex, because $\|x\|_1$ is convex.

— However, $f(x)$ is NOT differentiable, because $\|x\|_1$ is NOT.

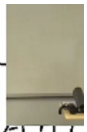
$\|x\|_1 = \sum_{i=1}^n |x_i|$



$$\nabla f(x^{(*)}) = 0$$

$$\Updownarrow$$

$$A^T(Ax^{(*)} - b) + \lambda x^{(*)} = 0 \iff \boxed{(A^TA + \lambda I)x^{(*)} = A^Tb}$$



— For any $\lambda > 0$, $A^TA + \lambda I$ is always invertible,

because $A^TA + \lambda I$ is alway SPD:

① $(A^TA + \lambda I)^T = A^TA + \lambda I$

② $x^T(A^TA + \lambda I)x = x^TA^TAx + \lambda x^Tx$

$= \|Ax\|_2^2 + \lambda \|x\|_2^2 > 0$ if $x \neq 0$.

— We can use steepest descent to find $x^{(*)}$.

- Kernel ridge regression.

choose a kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ s.t.

$$K(x, z) = \langle \phi(x), \phi(z) \rangle \quad \forall\, x, z \in \mathbb{R}^n.$$

$$\min_{C \in \mathbb{R}^N} \sum_{i=1}^N \left(\sum_{j=1}^N C_j K(x_i, x_j) - y_i \right)^2 + \lambda \sum_{i=1}^N \sum_{j=1}^N C_i C_j K(x_i, x_j)$$

$$\min_{C \in \mathbb{R}^N} \sum_{i=1}^N \left(\sum_{j=1}^N C_j K(x_i, x_j) - y_i \right)^2 + \lambda \sum_{i=1}^N \sum_{j=1}^N C_i C_j K(x_i, x_j)$$

$$\Updownarrow \quad \text{Let } K = [K(x_i, x_j)]_{i,j}$$

$$\min_{C \in \mathbb{R}^N} \frac{1}{2}\|KC - y\|_2^2 + \frac{\lambda}{2} C^TKC$$

$$\Updownarrow$$

$C^{(*)}$ is a solution $\iff K^T(KC^{(*)} - y) + \lambda KC^{(*)} = 0$

$$\iff (K^TK + \lambda K)C^{(*)} = K^Ty$$

Ex:
$f(x) = \frac{1}{2}x^TAx$
Then
$\nabla f(x) = \frac{1}{2}(A + A^T)x$

we have this we have this

§ 5.2.2. Neural Network Training

Given $\{x^{(i)}, y_i\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

We want to find a function f such that

$$f(x^{(i)}) \approx y_i \quad , \quad i = 1, 2, \cdots, m.$$

In linear regression, we choose f be an affine function

In ridge regression an affine function with a coefficien with small 2-norm

In LASSO regression an affine function with a coefficient is small 1-norm

In Kernel regression, a linear function in the feature space

the

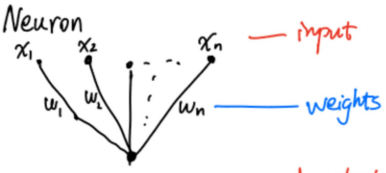
an

Neural network a function generated by

In kernel regression, a linear function in the feature space

In an Neural network model a function generated by artifical neural networks

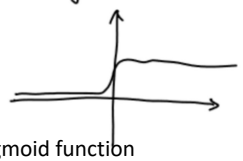
- Artificial Neuron



$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i\right) \quad \text{or} \quad y = \sigma(\langle w, x\rangle)$$

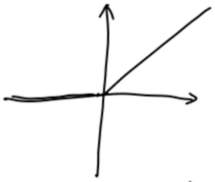
σ: activation

Activation function $\sigma: \mathbb{R} \to \mathbb{R}$



Sigmoid function

classical choice

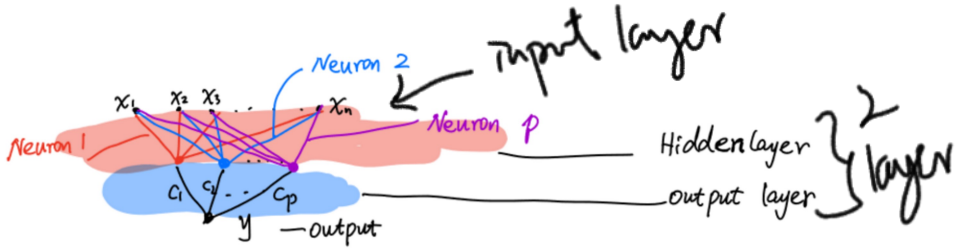


ReLU (popular recently)

... many others

- Neural Networks: Place many neurons together.

For simplicity, we first consider the following 2-layer neural network:



Let $w^{(j)} \in \mathbb{R}^n$ be the weights in neuron j, $j = 1, \cdots p$.

Denote $W = [w^{(1)} \cdots w^{(p)}] \in \mathbb{R}^{n \times p}$

Let $c = \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} \in \mathbb{R}^p$ be the weights in the output layer.

Then the input-output relation is

$$y = \sum_{j=1}^{p} c_j \sigma(\langle w^{(j)}, x\rangle)$$

$$= \langle c, \sigma(W^T x)\rangle \equiv f_{W,c}(x), \quad \text{where } x \in \mathbb{R}^n.$$

- Neural Network training:

— Find weights W, c to make the neural network function $f_{W,c}: \mathbb{R}^n \to \mathbb{R}$ fit the traing data $\{x^{(i)}, y_i\}_{i=1}^m$ the best.

— There are many different quantitative definition of "the best"

$f_{w,c}: \mathbb{R}^n \to \mathbb{R}$ fit the traing data $\{x^{(i)}, y_i\}_{i=1}^m$ the best.

— There are many different quantitative definition of "the best".

We may use the least squares error:

$$\min_{\substack{W \in \mathbb{R}^{n\times p} \\ c \in \mathbb{R}^p}} \sum_{i=1}^m \left(f_{w,c}(x^{(i)}) - y_i \right)^2$$

因為W有n*p dimension

c有p dimension

— Let $F_i(W,c) = \left(f_{w,c}(x^{(i)}) - y_i \right)^2$ and $F(W,c) = \sum_{i=1}^m F_i(W,c)$.

— Then $F, F_i : \mathbb{R}^{n\times p} \times \mathbb{R}^p \left(\equiv \mathbb{R}^{(n+1)p}\right) \to \mathbb{R}$ 合併可以咁寫

— Although F is NOT convex, we still apply gradient descent algorithm (or its variants) to train the neural network.

— A mysterious fact is that: despite of non-convexity, gradient-type algorithms always train a good neural network. This fact inspires the recent burst of research and applications of AI in the real world.

雖然係NN係non convex的，但係gradient descent依然可以穩到好近ans嘅答案

— We need to compute $\nabla F(W,p)$.

— Since F is a function on a Euclidean space $\mathbb{R}^{(n+1)p}$, we have

$$\nabla F(W,p) = \begin{bmatrix} \frac{\partial F}{\partial w^{(1)}}(W,c) \\ \vdots \\ \frac{\partial F}{\partial w^{(p)}}(W,c) \\ \frac{\partial F}{\partial c}(W,c) \end{bmatrix}$$

, where $\frac{\partial F}{\partial w^{(j)}}$ stands for the gradient of F w.r.t $w^{(j)}$ when other variables are fixed.

By direct calculation,

$$\frac{\partial F}{\partial w^{(j)}}(W,c) = \frac{\partial}{\partial w^{(j)}} \sum_{i=1}^m \left(f_{w,c}(x^{(i)}) - y_i \right)^2$$

$$= \sum_{i=1}^m \left[\frac{\partial}{\partial w^{(j)}} \left(f_{w,c}(x^{(i)}) - y_i \right)^2 \right]$$

• Let $g_1(t) = (t - y_i)^2$, $g_2(w^{(j)}) = f_{w,c}(x^{(i)})$ variable for w(j) only, other = fixed

So, $g_1: \mathbb{R} \to \mathbb{R}$, $g_2: \mathbb{R}^n \to \mathbb{R}$, and $(g_1 \circ g_2)(w^{(j)}) = \left(f_{w,c}(x^{(i)}) - y_i \right)^2$

By the chain rule,

$$\frac{\partial}{\partial w^{(j)}} \left(f_{w,c}(x^{(i)}) - y_i \right)^2 = g_1'(g_2(w^{(j)})) \cdot \nabla g_2(w^{(j)})$$

$$= 2 \left(f_{w,c}(x^{(i)}) - y_i \right) \cdot \frac{\partial}{\partial w^{(j)}} f_{w,c}(x^{(i)})$$

• Because $f_{w,c}(x^{(i)}) = \sum_{k=1}^p c_k \sigma(\langle w^{(k)}, x^{(i)} \rangle)$

$$= c_j \sigma(\langle w^{(j)}, x^{(i)} \rangle) + \sum_{k \neq j} c_k \sigma(\langle w^{(k)}, x^{(i)} \rangle)$$

constant of $w^{(j)}$ denoted by A.

$$= c_j \sigma(\langle w^{(j)}, x^{(i)} \rangle) + A$$

Let $g_3(w^{(j)}) = \langle w^{(j)}, x^{(i)} \rangle$ [So, $g_3: \mathbb{R}^n \to \mathbb{R}$)

[So, $g: \mathbb{R} \to \mathbb{R}$)

Let $g_3(w^{(j)}) = \langle w^{(j)}, x^{(i)} \rangle$ (So, $g_3 : \mathbb{R}^n \to \mathbb{R}$)

$g_4(t) = c_j \sigma(t) + A$ (So, $g_4 : \mathbb{R} \to \mathbb{R}$)

Obviously, $(g_4 \circ g_3)(w^{(j)}) = g_4(g_3(w^{(j)})) = c_j \sigma(\langle w^{(j)}, x^{(i)} \rangle) + A = f_{w,c}(x^{(i)})$

By the chain rule,

$$\frac{\partial}{\partial w^{(j)}} f_{w,c}(x^{(i)}) = \nabla (g_4 \circ g_3)(w^{(j)}) = g_4'(g_3(w^{(j)})) \cdot \nabla g_3(w^{(j)})$$

Chain rule

Direct calculation:

$g_4'(t) = c_j \sigma'(t)$, $\quad \nabla g_3(w^{(j)}) = x^{(i)}$

Thus, $\frac{\partial}{\partial w^{(j)}} f_{w,c}(x^{(i)}) = c_j \sigma'(\langle w^{(j)}, x^{(i)} \rangle)\, x^{(i)}$.

- Altogether,

$$\frac{\partial E}{\partial w^{(j)}}(W,c) = \sum_{i=1}^{m} \left[\frac{\partial}{\partial w^{(j)}} \left(f_{w,c}(x^{(i)}) - y_i \right)^2 \right]$$
$$= \sum_{i=1}^{m} 2 \left(f_{w,c}(x^{(i)}) - y_i \right) c_j \sigma'(\langle w^{(j)}, x^{(i)} \rangle)\, x^{(i)}$$
$$= 2 c_j \sum_{i=1}^{m} \left[\left(f_{w,c}(x^{(i)}) - y_i \right) \cdot \sigma'(\langle w^{(j)}, x^{(i)} \rangle) \right] x^{(i)}$$

- For $\frac{\partial E}{\partial c}$,

$$\frac{\partial E}{\partial c}(W,c) = \sum_{i=1}^{m} \frac{\partial}{\partial c} \left(f_{w,c}(x^{(i)}) - y_i \right)^2$$
$$= \sum_{i=1}^{m} 2 \left(f_{w,c}(x^{(i)}) - y_i \right) \cdot \frac{\partial}{\partial c} f_{w,c}(x^{(i)})$$

Since $\frac{\partial}{\partial c} f_{w,c}(x^{(i)}) = \frac{\partial}{\partial c} \left(\sum_{k=1}^{p} c_k \sigma(\langle w^{(k)}, x^{(i)} \rangle) \right)$

$$= \frac{\partial}{\partial c} \langle c, \sigma(W^T x^{(i)}) \rangle$$ (constant of c.)

$$= \sigma(W^T x^{(i)})$$

Thus, $\frac{\partial E}{\partial c}(W,c) = 2 \sum_{i=1}^{m} \left(f_{w,c}(x^{(i)}) - y_i \right) \cdot \sigma(W^T x^{(i)})$

The gradient descent for neural network training is

(k)

$$\begin{cases} w^{(j,k+1)} = w^{(j,k)} - \alpha_k \cdot 2 c_j^{(k)} \sum_{i=1}^{m} \left[\left(f_{W^{(k)},c^{(k)}}(x^{(i)}) - y_i \right) \cdot \sigma'(\langle w^{(j,k)}, x^{(i)} \rangle) \right] x^{(i)}, & j = 1, \dots, p \\ c^{(k+1)} = c^{(k)} - \alpha_k \cdot 2 \sum_{i=1}^{m} \left(f_{W^{(k)},c^{(k)}}(x^{(i)}) - y_i \right) \cdot \sigma\left((W^{(k)})^T x^{(i)} \right) \end{cases}$$

where $W^{(k)} = [w^{(1,k)}, \dots, w^{(p,k)}]$ and $c^{(k)}$ are the current weights.

In the computation, in addition to $x^{(i)}, y_i$, the other vectors are all output of the neural networks at different stages

— $f_{W^{(k)},c^{(k)}}(x^{(i)})$, output of the neural network in the end

— $\sigma\left((W^{(k)})^T x^{(i)} \right)$, output of the hidden neurons

— $\langle w^{(j,k)}, x^{(i)} \rangle$, output of the j-th hidden neuron before activation.

This algorithm is also known as Back Propagation.

besides these steps mentioned above, other operations are just vector operation

back propagation : we pass x_i in the NN and then use output to train the NN back

- Stochastic Gradient descent (SGD)

In big data application, $\{x^{(i)}, y_i\}_{i=1}^{m}$ with a huge m.

Therefore, the "$\sum_{i=1}^{m}$" in the gradient may be too expensive.

So, we don't use ALL training data in one batch. Instead, we

- Stochastic Gradient descent (SGD)

  In big data application, $\{x^{(i)}, y_i\}_{i=1}^m$ with a huge m.

  Therefore, the "$\sum_{i=1}^m$" in the gradient may be too expensive.

  So, we don't use ALL training data in one batch. Instead, we randomly sample some data $\{x^{(i)}, y_i\}_{i\in I}$, where $I \subset \{1,2,\cdots,m\}$ and $|I|$ is relatively small

  Then, we apply gradient descent to

$$\min_{W,C} \sum_{i\in I} (f_{W,C}(x^{(i)}) - y_i)^2.$$

  The resulting algorithm is known as SGD or mini-batch SGD.

  for $k = 1, 2, \cdots$

  randomly choose $I \subset \{1,\cdots,m\}$ with $|I|$ small.

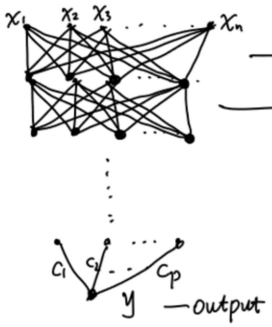
$$w^{(j,k+1)} = w^{(j,k)} - \alpha_k \cdot 2 c_j^{(k)} \sum_{i\in I} \left[\left(f_{W^{(k)},c^{(k)}}(x^{(i)}) - y_i \right) \cdot \sigma'\left(\langle w^{(j,k)}, x^{(i)} \rangle \right) \right] x^{(i)}, \quad j = 1, 2, \cdots, p.$$

$$c^{(k+1)} = c^{(k)} - \alpha_k \cdot 2 \sum_{i\in I} \left(f_{W^{(k)},c^{(k)}}(x^{(i)}) - y_i \right) \cdot \sigma\left((W^{(k)})^T x^{(i)} \right)$$

  end for

- Deep learning (deep neural network)

  Use many hidden layers



  Similar to one hidden layer case.

$$y = \langle c, \underbrace{\sigma\left(W^T \cdots \sigma\left((W^{(2)})^T \sigma\left((W^{(1)})^T x \right) \right) \right)}_{f_{W,c}(x)} \rangle$$

$$\min_{W,C} \sum_{i=1}^N (f_{W,c}(x^{(i)}) - y_i)^2$$

§ 6.2 Unconstrained Non-Smooth Convex Optimization

We consider

$$\min_{x\in\mathbb{R}^n} g(x)$$

where $g(x)$ is non-differentiable but convex.

$$\min_{W,C} \sum_{i=1}^{N} \left(f_{W,C}(x^{(i)}) - y_i \right)^2$$

§ 6.2 Unconstrained Non-Smooth Convex Optimization

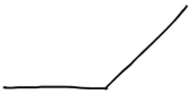
We consider

$$\min_{x \in \mathbb{R}^n} \; g(x)$$

where $g(x)$ is non-differentiable but convex.

Examples of non-differentiable convex functions.

- 1-norm (in LASSO and sparsity recovery)
- ReLU function (in deep learning)



- loss function $h(t)$ in "soft" SVM

§ 6.2.1. Sub-differential / Sub-gradient and Optimality

- To give an optimality condition, we need to extend differentiation to non-differentiable convex functions.

To this end, we first prove:

redefine gradient

- Theorem: Assume $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Then, for any given vector $x \in \mathbb{R}^n$, want to show these two are the same

$$\{\nabla f(x)\} = \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y-x \rangle \quad \forall y \in \mathbb{R}^n\}.$$

proof: Since, if f is convex and differentiable, $\nabla f(x)$ satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \quad \forall y \in \mathbb{R}^n,$$

we have

this definiton is only for convex

$$\{\nabla f(x)\} \subseteq \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y-x \rangle, \; \forall y \in \mathbb{R}^n\}$$

It remains to prove

$$\{\nabla f(x)\} = \left\{ u \;\middle|\; \lim_{\|y-x\| \to 0} \frac{f(y) - \left(f(x) + \langle u, y-x \rangle\right)}{\|y-x\|} = 0 \right\}$$

this is the definition of gradient before, for convex & non convex

then we restrict the case to convex case

and then we realize that the limit for ||y-x|| -> 0 is not generalizable to high dimensional case
since y can approach to x in different direction
=>> want to find another definition without limit

remain to prove this shit

remain to prove this shit

$$\{\nabla f(x)\} \supseteq \{u\in\mathbb{R}^n \mid f(y)\geq f(x)+\langle u, y-x\rangle,\ \forall y\in\mathbb{R}^n\}$$

To this end, let $u\in\mathbb{R}^n$ be satisfying $f(y)\geq f(x)+\langle u, y-x\rangle\ \forall y\in\mathbb{R}^n$.

Since f is convex, $f(2x-y)\geq f(x)+\langle \nabla f(x), x-y\rangle$

Therefore, $f(y)+f(2x-y)-2f(x)\geq \langle u-\nabla f(x), y-x\rangle \quad \forall y\in\mathbb{R}^n$. ---(1)

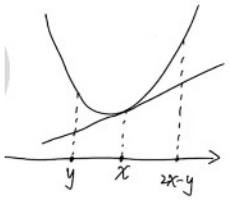
Similarly, $f(2x-y)\geq f(x)+\langle u, x-y\rangle$

$f(y)\geq f(x)+\langle \nabla f(x), y-x\rangle$ switch gradient(f) and u(from convexity)

$\Rightarrow f(y)+f(2x-y)-2f(x)\geq -\langle u-\nabla f(x), y-x\rangle \quad \forall y\in\mathbb{R}^n$. ---(2)

Since $f(2x-y)+f(y)-2f(x)=2\left(f(2x-y)+f(y)-f\left(\frac{1}{2}(2x-y)+\frac{1}{2}y\right)\right)\geq 0$,

combing (1)&(2) gives $|\langle u-\nabla f(x), y-x\rangle|\leq |f(y)+f(2x-y)-2f(x)|$ → this is >0 so abs sign can be omitted



from convexity and differentiability

Also, $f(y)=f(x)+\langle \nabla f(x), y-x\rangle+o(\|x-y\|)$

$f(2x-y)=f(x)+\langle \nabla f(x), x-y\rangle+o(\|x-y\|)$

$\Rightarrow f(y)+f(2x-y)-2f(x)=o(\|x-y\|)$,

i.e., $\lim\limits_{y\to x}\frac{|f(y)+f(2x-y)-2f(x)|}{\|x-y\|}=0$.

which implies

$$0\leq \lim_{y\to x}\frac{|\langle u-\nabla f(x), y-x\rangle|}{\|y-x\|}\leq \lim_{y\to x}\frac{|f(y)+f(2x-y)-2f(x)|}{\|x-y\|}=0$$

$$\Rightarrow \lim_{y\to x}\frac{|\langle u-\nabla f(x), y-x\rangle|}{\|y-x\|}=0.$$

Now, if we take $y=x+t(u-\nabla f(x))$ with $t\in\mathbb{R}$, choose y as a special path to approach 0

then $\lim\limits_{t\to 0}\frac{|\langle u-\nabla f(x), y-x\rangle|}{\|y-x\|}=\lim\limits_{t\to 0}\frac{t\|u-\nabla f(x)\|^2}{t\|u-\nabla f(x)\|}=\|u-\nabla f(x)\|$

$$\| \quad \lim_{y\to x}\frac{|\langle u-\nabla f(x), y-x\rangle|}{\|y-x\|}=0$$

Therefore $u=\nabla f(x)$. ⊠

- Therefore, we can use the set $\{u \mid f(y)\geq f(x)+\langle u, y-x\rangle, \forall y\in\mathbb{R}^n\}$ to define gradient of a convex differentiable function f.
  Notice that there is no limit in this equivalent definition. So, we can extend it to convex non-differentiable function.
- Sub-differential / sub-gradient:

- We call $\partial g(x)$ — sub-differential of g
  and elements in $\partial g(x)$ — sub-gradients of g.

Given a convex function $g: \mathbb{R}^n \to \mathbb{R}$ (differentiable or non-differentiable), its sub-differential at $x \in \mathbb{R}^n$, denoted by $\partial g(x)$, is defined by

$$\partial g(x) = \{ u \in \mathbb{R}^n \mid g(y) \geq g(x) + \langle u, y-x \rangle \quad \forall y \in \mathbb{R}^n \}.$$

Any element in $\partial g(x)$ is call a sub-gradient.

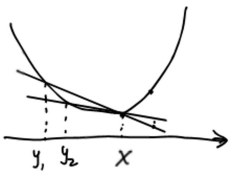
- From previous theorem, if g is differentiable at x, then
$$\partial g(x) = \{ \nabla f(x) \}.$$
- What if g is not differentiable at $x \in \mathbb{R}^n$?

  Theorem: If g is NOT differentiable at $x \in \mathbb{R}^n$, then $\partial g(x)$ contains more than one element.

  proof. We prove only the 1-D case, i.e., $g: \mathbb{R} \to \mathbb{R}$ is convex.

  Let $y_1 \leq y_2 < x$. Then, $\exists\, t \in (0,1]$ s.t.

  $$y_2 = t y_1 + (1-t)x = x + t(y_1 - x)$$

  Let $S(y) = \dfrac{g(y)-g(x)}{y-x}$.

  Since g is convex,

  $$\begin{aligned} S(y_2) - S(y_1) &= \frac{g(y_2)-g(x)}{y_2-x} - \frac{g(y_1)-g(x)}{y_1-x} \\ &= \frac{g(t y_1 + (1-t)x) - g(x)}{t y_1 + (1-t)x - x} - \frac{g(y_1)-g(x)}{y_1-x} \\ &\geq \frac{t g(y_1) + (1-t) g(x) - g(x)}{t(y_1-x)} - \frac{g(y_1)-g(x)}{y_1-x} \\ &= \frac{t(g(y_1)-g(x))}{t(y_1-x)} - \frac{(y_1)\, g(x)}{y_1-x} = 0 \end{aligned}$$

  (because $y_2 - x < 0$)

  Therefore, $S(y)$ is monotonically non-increasing on $[y_1, x)$.

  Furthermore, let $z > x$. Then, due to the convexity,

  (A) $\cdots\cdots$ $g(z) \geq g(x) + S(y)(z-x), \quad \forall\, y \in [y_1, x)$

  because otherwise letting $t = \dfrac{z-x}{z-y} \in (0,1)$ gives

  $$x = z + t(y-z) = t y + (1-t) z$$

  and

  $$\begin{aligned} (x) &> g(z) - \frac{g(y)-g(x)}{y-x}(z-x) \\ &= g(z) - \frac{g(y)-g(x)}{(t-1)(z-y)} \cdot t(z-y) = g(z) + \frac{t(g(y)-g(x))}{1-t} \end{aligned}$$

$\Rightarrow (1-t)g(x) > (1-t)g(z) + tg(y) - tg(x)$

$\Rightarrow g(x) = g(ty+(1-t)z) > (1-t)g(z) + tg(y)$.

Contradiction with convexity of g.

Eq. (A) implies $S(y) \le \frac{g(z)-g(x)}{z-x} \equiv S(z)$.

To sum up, $S(y)$ is monotonically non-increasing on $[y_1, x)$

and $S(y) \le S(z) \quad \forall\, z > x$.

Therefore, $\lim_{y \to x_-} S(y) \equiv D_- g(x)$ exists and

Similarly, $\lim_{z \to x_+} S(z) \equiv D_+ g(x)$ exisists.

Also, $S(y) \le D_- g(x) \le D_+ g(x) \le S(z) \quad \forall\, y<x$ and $z>x$.

Choose $u \in [D_- g(x), D_+ g(x)]$. Then:

Case $y<x$: $g(y) = g(x) + \left(\frac{g(y)-g(x)}{y-x}\right)(y-x)$

$= g(x) + S(y)(y-x) \ge (x) + u(y-x)$

Case $y>x$: $g(y) = g(x) + \left(\frac{g(y)-g(x)}{y-x}\right)(y-x)$

$= g(x) + S(y)(y-x) \ge (x) + u(y-x)$

Case $y=x$: $g(y) = g(x) + u(y-x)$

Thus, $\partial g(x) = [D_- g(x), D_+ g(x)]$, which is not empty.

When g is differentiable, $D_- g(x) = D_+ g(x)$, i.e., there is only one element in $\partial g(x)$.

When g is Not differentiable, $D_- g(x) < D_+ g(x)$, $\partial g(x)$ is an interval and contains infinitely many elements. ⊠.

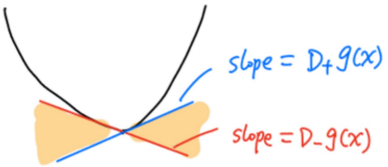
- From the proof, we see that: the one-sided derivative of a convex function $g: \mathbb{R} \to \mathbb{R}$ always exists. Also,

$$\partial g(x) = [D_- g(x), D_+ g(x)]. \qquad \forall x \in \mathbb{R}.$$

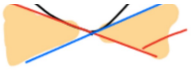
$$\partial g(x) = [D_- g(x), D_+ g(x)],$$

where $D_- g(x) = \lim_{y \to x_-} \frac{g(y)-g(x)}{y-x}$

$$D_+ g(x) = \lim_{y \to x_+} \frac{g(y)-g(x)}{y-x}$$



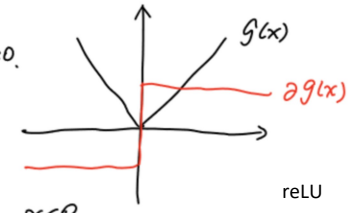
Example: $g(x) = |x| \quad x \in \mathbb{R}$.

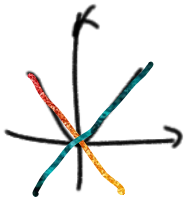


Example 1: $g(x) = |x| \quad x \in \mathbb{R}$.

$g(x)$ is NOT differentiable at only $x=0$.

$$\partial g(x) = \begin{cases} \{-1\} & \text{if } x<0 \\ [-1,1] & \text{if } x=0 \\ \{1\} & \text{if } x>0 \end{cases}$$





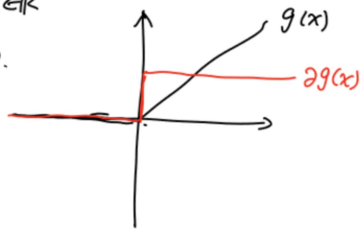
slope can only in $[-1,1]$

Example 2: $g(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } x \ge 0 \end{cases}, \quad x \in \mathbb{R}$

reLU

$g(x)$ is NOT differentiable at only $x=0$.

$$\partial g(x) = \begin{cases} \{0\} & \text{if } x<0 \\ [0,1] & \text{if } x=0 \\ \{1\} & \text{if } x>0 \end{cases}$$



Example 3: $g(x) = \|x\|_2, \quad x \in \mathbb{R}^n$.

$g(x)$ is differentiable at $x \neq 0$, and $\nabla g(x) = \frac{x}{\|x\|_2}$. if $x \neq 0$.

Therefore $\partial g(x) = \left\{\frac{x}{\|x\|_2}\right\}$ if $x \neq 0$.

Let us find $\partial g(0)$ by definition.

$$\partial g(0) = \{ u \in \mathbb{R}^n \mid g(y) \ge g(0) + \langle u, y \rangle, \ \forall y \in \mathbb{R}^n \}$$

$$= \{ u \in \mathbb{R}^n \mid \|y\|_2 \ge \langle u, y \rangle, \ \forall y \in \mathbb{R}^n \} \equiv S$$

呢到要做啲轉化先

Claim: $S = U \equiv \{ u \in \mathbb{R}^n \mid \|u\|_2 \le 1 \}$.

Proof. For any $u \in U$, $\|u\|_2 \le 1$, which implies

$$\langle u, y \rangle \le \|u\|_2 \|y\|_2 \le \|y\|_2, \ \forall y \in \mathbb{R}^n \implies u \in S.$$

First prove U is a subset of U

$\subset S$ ← So $U \subseteq S$. ← $u \in U$

For any $u \notin U$, i.e., $\|u\|_2 > 1$, we choose $y = u$.

Then $\langle u, y \rangle = \|u\|_2^2 = \|u\|_2 \|y\|_2 > \|y\|_2$

So $u \notin S$. Thus, $U^c \subseteq S^c \implies S \subseteq U$.

Then prove U^c in S^c

We can find a y s.t. S definition is not satisfied => u not in S

Therefore $S = U$. ☒

Altogether, $$\partial \|x\|_2 = \begin{cases} \left\{\frac{x}{\|x\|_2}\right\} & \text{if } x \neq 0 \\ \{u \in \mathbb{R}^n \mid \|u\|_2 \le 1\} & \text{if } x = 0. \end{cases}$$

- Sub-differential calculus rules:
  - $\partial(\alpha g) = \alpha \partial g \qquad \forall \alpha \in \mathbb{R}$.
  - $\partial(g_1 + g_2) = \partial g_1 + \partial g_2$
  - If $g(x) = f(Ax+b)$, then $\partial g(x) = A^T \partial f(Ax+b)$

Example 4: If $g(x) = \sum_{i=1}^n g_i(x_i)$, where $g_i : \mathbb{R} \to \mathbb{R}$ is convex,

Gi is simple calculation

then $\in \begin{bmatrix} u_1 \\ u_n \end{bmatrix}$

Example 4: If $g(x) = \sum_{i=1}^{n} g_i(x_i)$, where $g_i : \mathbb{R} \to \mathbb{R}$ is convex,

Gi is simple calculation

then $\partial g(x) = \left\{ \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \mid u_i \in \partial g_i(x_i) \right\}$

proof. Let $\tilde{g}_i(x) = g_i(x_i)$.

Then $\partial g(x) = \partial\left(\sum_{i=1}^{n} \tilde{g}_i(x)\right) = \sum_{i=1}^{n} \partial \tilde{g}_i(x)$.

Now lets show $\partial \tilde{g}_i(x) = \{a e_i \mid a \in \partial g_i(x_i)\}$. by direct calculation

$\partial \tilde{g}_i(x) = \{u \in \mathbb{R}^n \mid \tilde{g}_i(y) \geq \tilde{g}_i(x) + \langle u, y - x \rangle, \ \forall y \in \mathbb{R}^n\}$

$= \{u \in \mathbb{R}^n \mid g_i(y_i) \geq g_i(x_i) + u_i \cdot (y_i - x_i) + \sum_{j \neq i} u_j (y_j - x_j), \ \forall y \in \mathbb{R}^n\}$

- $\partial \tilde{g}_i(x) \supseteq \{u \in \mathbb{R}^n \mid u_i \in \partial g_i(x_i), \ u_j = 0 \ \forall j \neq i\} = \{a e_i \mid a \in \partial g_i(x_i)\}$
- If $u \neq a e_i$, where $a \in \partial g_i(x_i)$, and $u \in \partial \tilde{g}_i(x)$, then,
  - If $u_j = 0 \ \forall j \neq i$, then

$g_i(y_i) \geq g_i(x_i) + u_i(y_i - x_i) + \sum_{j \neq i} u_j (y_j - x_j) = g_i(x_i) + u_i(y_i - x_i) \quad \forall y_i \in \mathbb{R}$

$\Rightarrow u_i \in \partial g_i(x_i) \Rightarrow u = a e_i, \ a \in \partial g_i(x_i)$

  - Therefore, $\exists j \neq i$ s.t $u_j \neq 0$. Choose $y = (c|u_j| + x_j) e_j$ with $c \in \mathbb{R}$

$g_i(0) \geq g_i(x_i) + u_i(y_i - x_i) + \sum_{j \neq i} u_j (y_j - x_j)$

$= g_i(x_i) + c|u_j|^2 \longrightarrow +\infty$ as $c \to +\infty$. contradiction.

Thus, $\partial \tilde{g}_i(x) = \{a e_i \mid a \in \partial g_i(x_i)\}$

We obtain $\partial g(x) = \{\sum_{i=1}^{n} a_i e_i \mid a_i \in \partial g_i(x_i)\}$

$= \left\{ \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \mid u_i \in \partial g_i(x_i) \right\}$

Skipped by prof

Example 6: $g(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$, where $\lambda > 0$.

$\partial g(x) = A^T(Ax - b) + \lambda \, \partial \|x\|_1$

$$= \left\{ \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \mid u_i \in \partial g_i(x_i) \right\}$$

Example 5: $\partial \|x\|_1 = \left\{ \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \mid u_i \in \partial |x_i| \right\}$, because $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

Example 6: $\partial \|Dx\|_1 = \left\{ D^T \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \mid u_i \in \partial |(Dx)_i|_1 \right\}$, where $D \in \mathbb{R}^{m \times n}$.

$\partial |(Dx)_i|_1$

- Optimality:

Fermat's Lemma: Assume $g: \mathbb{R}^n \to \mathbb{R}$ is convex. Then

$$x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff 0 \in \partial g(x^{(*)})$$

Proof. $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff g(x) \geq g(x^{(*)}) \quad \forall\, x \in \mathbb{R}^n$

$\iff g(x) \geq g(x^{(*)}) + \langle 0, x - x^{(*)} \rangle \quad \forall\, x \in \mathbb{R}^n$

$\iff 0 \in \partial g(x^{(*)})$ (W) ⊠

Example 1: If g is convex and differentiable, then $\partial g(x) = \{\nabla g(x)\}$.

Therefore, $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff \nabla g(x^{(*)}) = 0$.

Example 2: $g(x) = \|x\|_1$

$= \left\{ \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \mid u_i \in [-1, 1] \quad \forall i \right\}$

If x is non zero, some coordinates Are also non zero, u_i must be Non zero, and subgrad(g(x)) will be Non zero

$0 \in \partial \|x^{(*)}\|_1 \iff 0 \in \left\{ \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \mid u_i \in \partial |x_i^{(*)}| \right\} \iff x^{(*)} = 0.$

0 is the only global minimizer of g(x)

Therefore, $g(x) = \|x\|_1$ is minimized at 0.

Example 3: $g(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$.

$$\partial g(x) = A^T(Ax - b) + \lambda\, \partial \|x\|_1$$

Therefore, $x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x) \iff 0 \in A^T(Ax^{(*)} - b) + \lambda\, \partial \|x^{(*)}\|_1$

$\iff \frac{1}{\lambda} A^T(b - Ax^{(*)}) \in \partial \|x^{(*)}\|_1$

Therefore, $\left[\frac{1}{\lambda} A^T(b - Ax^{(*)})\right]_i \in [-1, 1]$.

If $\left|\left[\frac{1}{\lambda} A^T(b - Ax^{(*)})\right]_i\right| < 1$, then $[x^{(*)}]_i = 0$.

If $\left[\frac{1}{\lambda} A^T(b - Ax^{(*)})\right]_i = 1$, then $x_i^{(*)} \geq 0$

$\left[\frac{1}{\lambda} A^T(b - Ax^{(*)})\right]_i = -1$, then $x_i^{(*)} \leq 0$

$\left|\left[\frac{1}{\lambda} A^T(b - Ax^{(*)})\right]_i\right| < 1$, then $x_i^{(*)} = 0$

($\Rightarrow$ the $x^{(*)}$ contains some 0 components)

When lambda is very large, 1/lambda*A^T(b-Ax$^{(*)}$ is small

then $\exists$ many i, s.t. $\left|\left[\frac{1}{\lambda}A^T(b-Ax^{(k)})\right]_i\right| < 1$

i.e., there are many 0's in $x^{(k)}$ vector

$\Rightarrow$ $x^{(k)}$ is sparse.

Lasso will give us sparse solution

§ 6.2.2. Sub-Gradient Descent

Consider $\min_{x\in\mathbb{R}^n} g(x)$, where g is convex but non-differentiable.

- Given $x^{(k)}\in\mathbb{R}^n$, let $u^{(k)}\in\partial g(x^{(k)})$. Then, by convexity,

$$g(x^{(k+1)}) \geq g(x^{(k)}) + \langle u^{(k)}, x^{(k+1)}-x^{(k)}\rangle$$

$$\Rightarrow g(x^{(k+1)}) - g(x^{(k)}) \geq \langle u^{(k)}, x^{(k+1)}-x^{(k)}\rangle$$

The right-hand side is minimized if

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k)}, \quad \text{where } \alpha_k > 0.$$

This is called sub-gradient descent. (or called forward sub-gradient)

In order to make sub gradient work
We need to choose small step size
To make:

- However, the right-hand side is only a lower bound, and there might be a gap as big as $O(\|x^{(k+1)}-x^{(k)}\|)$, which makes $g(x^{(k+1)}) > g(x^{(k)})$ for any $\alpha_k > 0$.

Example: $g(x) = |x|$., $x^{(k)} = 0$. Choose $u^{(k)} = -1 \in \partial g(x^{(k)})$.

Then $x^{(k+1)} = x^{(k)} - \alpha_k(-1) = \alpha_k$.

Left hand side $= g(x^{(k+1)}) - g(x^{(k)}) = \alpha_k$

right hand side $= \langle u^{(k)}, x^{(k+1)}-x^{(k)}\rangle = \langle -1, \alpha_k\rangle = -\alpha_k$.

the gap is

$$g(x^{(k+1)}) - g(x^{(k)}) - \langle u^{(k)}, x^{(k+1)}-x^{(k)}\rangle = 2\alpha_k$$

which is $2\|x^{(k+1)}-x^{(k)}\|. \sim O(\|x^{(k+1)}-x^{(k)}\|)$,

Consequently,

$$g(x^{(k+1)}) - g(x^{(k)}) = \langle u^{(k)}, x^{(k+1)}-x^{(k)}\rangle + 2\alpha_k$$

$$= -\alpha_k \|u^{(k)}\|^2 + 2\alpha_k = \alpha_k > 0.$$

In general, to make sub-gradient work, we have to choose

$$\alpha_k \to 0, \text{ as } k\to+\infty.$$

In summary, sub-gradient converges very slowly.

- We may use backward sub-gradient descent.

$$g(x^{(k+1)}) = g(x^{(k)}) + \langle \nabla g(x^{(k)}), x^{(k+1)}-x^{(k)}\rangle + o(\|x^{(k+1)}-x^{(k)}\|)$$

↑ little o.

descent.

- What is the step size α_k? Will sub-Gradient work?
  - Gradient descent work because we choose a small α_k s.t.

$$g(x^{(k+1)}) = g(x^{(k)}) + \langle \nabla g(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + o(\|x^{(k+1)} - x^{(k)}\|)$$

↑ little o.

  - but in the non-smooth case

$$g(x^{(k+1)}) = g(x^{(k)}) + \langle u, x^{(k+1)} - x^{(k)} \rangle + O(\|x^{(k+1)} - x^{(k)}\|)$$

$u \in \partial g(x^{(k)})$

↑ Can be a big O even α_k is very small.

Let $x^{(k)} \in \mathbb{R}^n$ be current iteration, and $x^{(k+1)} \in \mathbb{R}^n$ be the next iteration.

Instead lower bound of $g(x^{(k+1)}) - g(x^{(k)})$, we estimate an upper bound of $g(x^{(k+1)}) - g(x^{(k)})$, i.e.,

$$g(x^{(k+1)}) - g(x^{(k)}) \leq \cdots\cdots$$

So we need $g(x^{(k)}) \geq g(x^{(k+1)}) \cdots\cdots$

The idea is to expand g at $x^{(k+1)}$. For this, we need sub-gradient at $x^{(k+1)}$. Let $u^{(k+1)} \in \partial g(x^{(k+1)})$. Then, the convexity implies

$$g(x^{(k)}) \geq g(x^{(k+1)}) + \langle u^{(k+1)}, x^{(k)} - x^{(k+1)} \rangle,$$

i.e., $$g(x^{(k+1)}) - g(x^{(k)}) \leq \langle u^{(k+1)}, x^{(k+1)} - x^{(k)} \rangle \quad \cdots\cdots (A)$$

The right hand side is minimized when $x^{(k+1)} - x^{(k)} = -\alpha_k u^{(k+1)}$, i.e.,

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k+1)}, \quad \alpha_k \geq 0 \qquad \cdots\cdots (C)$$

This is called backward sub-gradient descent. 以前是 $\alpha_k u^{(k)}$, 宜多用 backward.

Then, (A) implies

$$g(x^{(k+1)}) - g(x^{(k)}) \leq -\alpha_k \|u^{(k+1)}\|^2 \qquad \cdots\cdots (B)$$

Thus,

(i). if $\alpha_k > 0$, then $g(x^{(k+1)}) \leq g(x^{(k)})$, i.e., the backward sub-gradient descent gives a monotonical non-increasing $\{g(x^{(k)})\}_{k=0}^{+\infty}$. Moreover, if $\exists$ a solution of $\min_{x \in \mathbb{R}^n} g(x)$, (i.e. $\exists\, x^{(*)} = \arg\min_{x \in \mathbb{R}^n} g(x)$), then $\lim_{k \to +\infty} g(x^{(k)}) = C$ exists.

(ii). if $\alpha_k \geq \alpha > 0$, then summing (B) over k gives

$$g(x^{(k+1)}) - g(x^{(0)}) \leq -\sum_{k=0}^{K} \alpha_k \|u^{(k+1)}\|^2 \leq -\alpha \sum_{k=0}^{K} \|u^{(k+1)}\|^2$$

Sending $K \to +\infty$,

$$C - g(x^{(0)}) \leq -\alpha \sum_{k=1}^{\infty} \|u^{(k)}\|^2, \quad \text{i.e.,}$$

$$\sum_{k=1}^{\infty} \|u^{(k)}\|^2 \leq \frac{g(x^{(0)}) - C}{\alpha} < +\infty$$

Therefore, $\lim_{k \to +\infty} \|u^{(k)}\| = 0$, where $u^{(k)} \in \partial g(x^{(k)})$.

If $\{x^{(k)}\}_{k=1}^{+\infty}$ is bounded, i.e., $\exists M>0$ s.t. $\|x^{(k)}\|_2 \le M \ \forall k$,

then $g(x^{(*)}) \ge g(x^{(k)}) + \langle u^{(k)}, x^{(k)} - x^{(*)} \rangle \ge g(x^{(*)}) - \|u^{(k)}\| \, \|x^{(k)} - x^{(*)}\|$

$\ge g(x^{(k)}) - (M + \|x^{(*)}\|_2) \, \|u^{(k)}\|$

Sending $k \to +\infty$, we obtain

$\min_{x \in \mathbb{R}^n} g(x) = g(x^{(*)}) \ge C$, which implies $C = \min_{x \in \mathbb{R}^n} g(x)$.

Therefore, $\lim_{k \to +\infty} g(x^{(k)}) = \min_{x \in \mathbb{R}^n} g(x)$.

i.e., backward sub-gradient converges to a global minimum as long as $\alpha_k \ge \alpha > 0$. (用大過 upward bound 的都 converge)

Theorem: Consider $\min_{x \in \mathbb{R}^n} g(x)$, where $g: \mathbb{R}^n \to \mathbb{R}$ is convex. Let $\{x^{(k)}\}_k$ be generated by

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k+1)}, \quad \text{where } u^{(k+1)} \in \partial g(x^{(k+1)})$$

(一個走的固定步長 size)

Then, if $\alpha_k \ge \alpha > 0 \ \forall k$ and $\min_{x \in \mathbb{R}^n} g(x)$ has a solution, then,

(i) $\{g(x^{(k)})\}_k$ is monotonically non-increasing;

(ii) $\lim_{k \to +\infty} \|u^{(k)}\|_2 = 0$;

(iii) If $\|x^{(k)}\|_2 \le M \ \forall k$, then $\lim_{k \to +\infty} g(x^{(k)}) = \min_{x \in \mathbb{R}^n} g(x)$.

(兩點入之割的條件)

- However, given $x^{(k)} \in \mathbb{R}^n$, we cannot obtain $x^{(k+1)}$ directly by the iteration:

$$x^{(k+1)} = x^{(k)} - \alpha_k u^{(k+1)}, \quad \text{where } u^{(k+1)} \in \partial g(x^{(k+1)})$$

($x^{(k+1)}$ depends on $x^{(k+1)}$)

We need to solve the above equation, which is equivalent to

$x^{(k+1)} \in x^{(k)} - \alpha_k \partial g(x^{(k+1)})$ ← 因為有一個 set 的關係.

$\Updownarrow$

$0 \in x^{(k+1)} - x^{(k)} + \alpha_k \partial g(x^{(k+1)})$

$\Updownarrow$ ← Fermat's lemma

Guess its "subintegral"

$0 \in \partial \left(\frac{1}{2} \|x - x^{(k)}\|_2^2 + \alpha_k g(x) \right) \Big|_{x = x^{(k+1)}}$

$\Downarrow$ Fermat's lemma.

$x^{(k+1)} \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - x^{(k)}\|_2^2 + \alpha_k g(x)$

因為有 $\|x - x^{(k)}\|_2^2$, 表示 $x^{(k+1)}$ is a proximity of $x^{(k)}$.

再加, $x^{(k+1)}$ has a small effect $g(x)$

$\Rightarrow$ called proximal algorithm.

$$x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x), \quad \text{where } F_{\alpha_k}(x) = \frac{1}{2}\|x-x^{(k)}\|_2^2 + \alpha_k g(x)$$

- F_{α_k} is continuous (because all convex function is continuous).
- F_{α_k} is coercive if $\min_{x\in\mathbb{R}^n} g(x)$ exists a solution.

Therefore, $\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x)$ exists at least a solution.

- Furthermore, F_{α_k} is strictly convex.

Therefore, $\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x)$ has a unique solution.

Altogether, $x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} F_{\alpha_k}(x)$ is well-defined.

Thus, backward sub-gradient descent is rewritten as

$$x^{(k+1)} = \arg\min_{x\in\mathbb{R}^n} \left(\frac{1}{2}\|x-x^{(k)}\|_2^2 + \alpha_k g(x)\right), \quad k=0,1,2,\cdots.$$

- Proximity Operator

Definition: Let $g:\mathbb{R}^n\to\mathbb{R}$ be convex. Let $\lambda>0$ be a parameter. The proximity operator $\text{prox}_{\lambda g}:\mathbb{R}^n\to\mathbb{R}^n$ of g is defined by

$$\text{prox}_{\lambda g}(y) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \lambda g(x), \quad \forall y\in\mathbb{R}^n \qquad \boxtimes$$

- Then, the backward sub-gradient descent is

$$x^{(k+1)} = \text{prox}_{\alpha_k g}\left(x^{(k)}\right), \quad k=0,1,2,\cdots.$$

For this reason, the backward sub-gradient descent is also called the proximal algorithm.

- Example 1: $g(x) = \|x\|_1$

Then, $x \equiv \text{prox}_{\lambda g}(y) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \lambda\|x\|_1$

$$\Updownarrow$$

$$0 \in x - y + \lambda\,\partial\|x\|_1$$

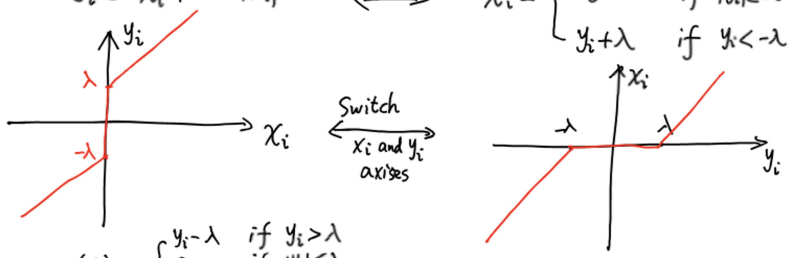
$$\Updownarrow$$

$$0 \in x_i - y_i + \lambda\,\partial|x_i|, \quad i=1,2,\cdots,n.$$

(Recall $\partial\|x\|_1 = \left\{\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \mid u_i \in \partial|x_i|\right\}$)

$$\Updownarrow$$

$$y_i \in x_i + \lambda \partial |x_i| \quad \Longleftrightarrow \quad x_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } |y_i| \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$



We define $T_\lambda(y_i) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } |y_i| \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$

Then, $\text{prox}_{\lambda \|\cdot\|_1}(y) = T_\lambda(y) \equiv \begin{bmatrix} T_\lambda(y_1) \\ T_\lambda(y_2) \\ \vdots \\ T_\lambda(y_n) \end{bmatrix}$ — called soft-thresholding operator.

The proximal algorithm for $\min_{x \in \mathbb{R}^n} \|x\|_1$ gives

$$x^{(k+1)} = T_{\alpha_k}(x^{(k)}) \quad , \quad k = 0, 1, 2, \cdots$$

At each iteration, components of $x^{(k)}$ are shrinked by α_k until it becomes 0. Therefore, $x^{(k)} \to 0 \equiv \arg\min_{x \in \mathbb{R}^n} \|x\|_1$. 囧

- Example 2: $g(x) = \|x\|_2^2$

  Then $x \equiv \text{prox}_{\lambda g}(y) = \arg\min_{x \in \mathbb{R}^n} \left(\frac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_2^2 \right)$

  $$\Updownarrow$$

  $$x - y + 2\lambda x = 0$$

  $$\Updownarrow$$

  $$x = \frac{1}{1+2\lambda} y$$

  The proximal algorithm for $\min_{x \in \mathbb{R}^n} \|x\|_2^2$ is

  $$x^{(k+1)} = \frac{1}{2 + 2\alpha_k} x^{(k)},$$

  which obviously converges to 0, the arg min of $\|x\|_2^2$.

- Example 3: $g(x) = \|x\|_2$

  Then $x \equiv \text{prox}_{\lambda g}(y) = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_2$

  $$\Updownarrow$$

  $$0 \in x - y + \lambda \partial \|x\|_2$$

  $$\Updownarrow$$

  $$y \in x + \lambda \partial \|x\|_2 \quad \cdots\cdots \quad (**)$$

$$\Updownarrow$$

$$0 \in x - y + \lambda \partial \|x\|_2$$

$$\Updownarrow$$

$$y \in x + \lambda \partial \|x\|_2 \quad \cdots\cdots \quad (**)$$

Let us find x.

If $x \neq 0$, then $y = x + \lambda \frac{x}{\|x\|_2} \Rightarrow x = cy$ with $c = \frac{1}{1+\frac{\lambda}{\|x\|_2}} \geq 0$.

If $x = 0$, then $\Rightarrow \quad x = cy$ with $c = 0$.

In any cases, the solution $x = cy$ with $c \geq 0$.

So the orginal problem $\Leftrightarrow \min_{c \geq 0} \frac{1}{2}(1-c)^2 \|y\|_2^2 + \lambda |c| \|y\|_2$

If $\|y\|_2 = 0$, obviously $c = 0$.

If $\|y\|_2 > 0$, then $\min_{c \geq 0} \frac{\|y\|_2}{2}(1-c)^2 + \lambda |c|$

$\Leftrightarrow \min_{c \geq 0} \frac{1}{2}(c-1)^2 + \frac{\lambda}{\|y\|_2}|c|$

Since $T_{\frac{\lambda}{\|y\|_2}}(1) \geq 0$ is the solution of $\min_{c \in \mathbb{R}} \frac{1}{2}(c-1)^2 + \frac{\lambda}{\|y\|_2}|c|$,

it is also the solution $\min_{c \geq 0} \frac{1}{2}(c-1)^2 + \frac{\lambda}{\|y\|_2}|c|$.

Therefore, $c = T_{\frac{\lambda}{\|y\|_2}}(1) = \begin{cases} 1 - \frac{\lambda}{\|y\|_2} & \text{if} \quad 1 \geq \frac{\lambda}{\|y\|_2}. \\ 0 & \text{if} \quad 1 \leq \frac{\lambda}{\|y\|_2}. \end{cases}$

$$= \begin{cases} \frac{\|y\|_2 - \lambda}{\|y\|_2} & \text{if} \quad \|y\|_2 \geq \lambda \\ 0 & \text{if} \quad \|y\|_2 \leq \lambda. \end{cases}$$

Finally, $\text{prox}_{\lambda g}(y) = x = \begin{cases} \frac{\|y\|_2 - \lambda}{\|y\|_2} y & \text{if} \quad \|y\|_2 \geq \lambda \\ 0 & \text{if} \quad \|y\|_2 \leq \lambda. \end{cases}$

The proximal algorithm for $\min_{x \in \mathbb{R}^n} \|x\|_2$ is

$$x^{(k+1)} = \begin{cases} \frac{\|x^{(k)}\|_2 - \alpha_k}{\|x^{(k)}\|_2} x^{(k)} & \text{if} \quad \|x^{(k)}\|_2 \geq \alpha_k \\ 0 & \text{if} \quad \|x^{(k)}\|_2 \leq \alpha_k \end{cases},$$

which obvious converges to 0.

- In general, it is difficult to find $\text{prox}_{\lambda g}(\cdot)$ for a given convex function g. Therefore, the proximal algorithm (Backward sub-gradient descent) is NOT practical.

function g. Therefore, the proximal algorithm (Backward sub-gradient descent) is NOT practical.

§ 6.2.3. Case Study: LASSO regression

LASSO regression model:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \qquad \text{(LASSO)}$$

↑ fit the given data. ↑ promote sparsity of the solution

- The objective function is convex.
- The objective function is continuous

The objective function is coercive, because

when $\|x\|_2 \to +\infty$, $\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \geq \lambda\|x\|_1 \geq \lambda\|x\|_2 \to +\infty$.

Therefore, there exists at least one solution.

- The uniqueness of the solution is NOT guaranteed, unless A satisfies some assumptions (e.g. A is invertible)
- The solution of (LASSO) is **sparse.**

To see this, let $x\in\mathbb{R}^n$ be a solution. Then

$$0 \in A^T(Ax-b) + \lambda\,\partial\|x\|_1$$

$$\Updownarrow$$

$$0 \in \alpha A^T(Ax-b) + \lambda\alpha\,\partial\|x\|_1 \qquad \forall \alpha>0$$

$$\Updownarrow$$

$$0 \in x - (x - \alpha A^T(Ax-b)) + \lambda\alpha\,\partial\|x\|_1 \qquad \forall \alpha>0$$

$$\Updownarrow$$

$$x = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \lambda\alpha\,\partial\|x\|_1, \quad \text{where } y = x-\alpha A^T(Ax-b)$$

$$\Updownarrow$$

$$x = \text{prox}_{\lambda\alpha\|\cdot\|_1}(x - \alpha A^T(Ax-b))$$
$$\Updownarrow$$
$$x = T_{\lambda\alpha}(x - \alpha A^T(Ax-b))$$

Recall that $T_{\lambda\alpha}(y)$ set y_i to 0 if $|y_i| \le \lambda\alpha$.

Therefore, x is a sparse if λ is large enough.

- Numerical solver for (LASSO).

The objective function in (LASSO) is NON-SMOOTH.

The gradient descent can not be applied.

But if the backward sub-gradient (proximal algorithm) is applied,

then we need the proximity operator, which solves

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \alpha\left(\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1\right)$$

This is the same difficulty as (LASSO).

We will use a "mixed" forward and backward (sub-)gradient algorithm.

Let $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ — convex and smooth

$g(x) = \|x\|_1$ — Convex but non-smooth

$$(\text{LASSO}) \iff \min_{x\in\mathbb{R}^n} f(x) + \lambda g(x)$$

$f(x)$: forward gradient; $\lambda g(x)$: backward sub-gradient

Forward use $x^{(k)}$, backward use $x^{(k+1)}$

We use the algorithm

$$x^{(k+1)} = x^{(k)} - \alpha_k\left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right), \quad \text{with } u^{(k+1)} \in \partial g(x^{(k+1)})$$

It is rewritten as

$$x^{(k+1)} \in x^{(k)} - \alpha_k \nabla f(x^{(k)}) - \alpha_k \lambda \partial g(x^{(k+1)})$$
$$0 \in x^{(k+1)} - (x^{(k)} - \alpha_k \nabla f(x^{(k)})) + \alpha_k \lambda \partial g(x^{(k+1)}) \quad \cdots (A)$$

Recall that

$$x = \text{prox}_{\beta g}(y) \iff x = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|x-y\|_2^2 + \beta g(x)$$
$$\iff 0 \in x - y + \beta\partial g(x) \quad \cdots (B)$$

By comparing (A) and (B), we obtain

$$x^{(k+1)} = \text{prox}_{\alpha_k \lambda g}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)$$

Since $\nabla f(x) = \nabla\left(\frac{1}{2}\|Ax-b\|_2^2\right) = A^T(Ax-b)$

$$\text{prox}_{\alpha_k \lambda g}(x) = T_{\alpha_k \lambda}(x)$$

the algorithm is

$$\boxed{x^{(k+1)} = T_{\alpha_k \lambda}\left(x^{(k)} - \alpha_k A^T(Ax^{(k)}-b)\right)}$$

This algorithm is known as iterative soft-thresholding algorithm

- From the derivation, for a generic optimization

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x),$$

where $f(x)$ is convex and smooth

$g(x)$ is convex but non-smooth,

we can use the forward-backward splitting (FBS) algorithm

$$x^{(k+1)} = x^{(k)} - \alpha_k\left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right), \quad u^{(k+1)} \in \partial g(x^{(k+1)})$$

which is the same as

$$x^{(k+1)} = \text{prox}_{\lambda \alpha_k g}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)$$

This is also known as proximal gradient algorithm.

- Convergence

Since f is convex and smooth,

(C) $\cdots f(x^{(k+1)}) = f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{1}{2}\langle \nabla^2 f(\tilde{x}^{(k)})(x^{(k+1)} - x^{(k)}), x^{(k+1)} - x^{(k)} \rangle$

(1) Since g is convex,

$$g(x^{(k)}) \geq g(x^{(k+1)}) + \langle u^{(k+1)}, x^{(k)} - x^{(k+1)} \rangle, \quad u^{(k+1)} \in \partial g(x^{(k+1)})$$

which is equivalent to

(D) $- - - g(x^{(k+1)}) \leq g(x^{(k)}) + \langle u^{(k+1)}, x^{(k+1)} - x^{(k)} \rangle$

(C) + λ(D) gives

$$\left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right]$$

$$\le \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] + \langle \nabla f(x^{(k)}) + \lambda u^{(k+1)},\ x^{(k+1)} - x^{(k)} \rangle + \frac{1}{2}\langle \nabla^2 f(\tilde{x}^{(k)})(x^{(k+1)} - x^{(k)}),\ x^{(k+1)} - x^{(k)} \rangle$$

Assuming $\langle \nabla^2 f(x)\, u,\ u \rangle \le M \|u\|_2^2 \quad \forall\, x, u \in \mathbb{R}^n$.

$$\le \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] - \alpha_k \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 + \frac{M\alpha_k^2}{2}\|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2$$

$$= \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] - \alpha_k\left(1 - \frac{M\alpha_k}{2}\right)\|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2$$

Assume $0 < \alpha_\ell \le \alpha_k \le \alpha_u < \frac{2}{M}$ for some $\alpha_\ell, \alpha_u \in \mathbb{R}$.

Then, $\exists\ \beta$ s.t. $\alpha_k\left(1 - \frac{M}{2}\alpha_k\right) \ge \beta > 0$.

So, $\left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right] - \left[f(x^{(k)}) + \lambda g(x^{(k)})\right] \le -\beta \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 \cdots (C)$

$\left\{f(x^{(k)}) + \lambda g(x^{(k)})\right\}_k$ is a monotonically non-increasing sequence.

(2) Assume $\min_{x \in \mathbb{R}^n} f(x) + \lambda g(x)$ has a solution, i.e.,

$\exists\, x^{(k)}$, s.t. $f(x^{(k)}) + \lambda g(x^{(k)}) = \min_{x \in \mathbb{R}^n} f(x) + \lambda g(x)$.

Then, $\lim_{k \to \infty} f(x^{(k)}) + \lambda g(x^{(k)}) = C$ exists.

Furthermore, summing (C) over k, we obtain

$$\left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right] - \left[f(x^{(0)}) + \lambda g(x^{(0)})\right] \le -\beta \sum_{k=0}^{K} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2$$

So $\sum_{k=0}^{K} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 \le \frac{1}{\beta}\left[\left(f(x^{(0)}) + \lambda g(x^{(0)})\right) - \left[f(x^{(k+1)}) + \lambda g(x^{(k+1)})\right]\right]$

$k \to +\infty$ gives $\sum_{k=0}^{+\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2^2 < +\infty$

Consequently, $\lim_{k \to +\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2 = 0$

Since $\alpha_k\left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right) = x^{(k)} - x^{(k+1)}$

$$0 \le \lim_{k \to +\infty} \|x^{(k)} - x^{(k+1)}\|_2 = \lim_{k \to +\infty} \|\alpha_k\left(\nabla f(x^{(k)}) + \lambda u^{(k+1)}\right)\|_2 \le \alpha_u \cdot \lim_{k \to +\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2 = 0,$$

which implies $\lim_{k \to +\infty} \|x^{(k)} - x^{(k+1)}\|_2 = 0$.

This together with the continuity of $\nabla f(x)$ implies $0 = \lim_{k \to +\infty} \|\nabla f(x^{(k)}) - \nabla f(x^{(k+1)})\|_2$

Therefore,

$$0 \le \lim_{k \to +\infty} \|\nabla f(x^{(k+1)}) + \lambda u^{(k+1)}\|_2 \le \lim_{k \to +\infty}\left(\|\nabla f(x^{(k)}) + \lambda u^{(k+1)}\|_2 + \|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\|_2\right) = 0,$$

i.e., $\lim_{k \to +\infty} \|\nabla f(x^{(k)}) + \lambda u^{(k)}\|_2 = 0$

③ Moreover,

$$\left[f(x^{(*)})+\lambda g(x^{(*)})\right] \geq \left[f(x^{(k)})+\lambda g(x^{(k)})\right] + \langle \nabla f(x^{(k)})+\lambda u^{(k)},\ x^{(*)}-x^{(k)}\rangle$$
$$\geq \left[f(x^{(k)})+\lambda g(x^{(k)})\right] - \|\nabla f(x^{(k)})+\lambda u^{(k)}\|_2\left(\|x^{(*)}\|_2+\|x^{(k)}\|_2\right)$$

Assume $\exists\ B>0$ s.t. $\|x^{(k)}\|_2 \leq B \quad \forall\ k$.

Then, sending $k\to+\infty$, we obtain

$$\min_{x\in\mathbb{R}^n} f(x)+\lambda g(x) \geq C. \quad \left(\text{Obviously, } C \geq \min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)\right).$$

Thus, $\boxed{\lim_{k\to+\infty} f(x^{(k)})+\lambda g(x^{(k)}) = \min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)}$

To sum up,

Theorem: Consider $\min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)$, where $\lambda>0$, f, g are convex.

Assume: ① $\langle \nabla^2 f(x)u, u\rangle \leq M\|u\|_2^2$, $\forall\ x, u\in\mathbb{R}^n$. (i.e., f is smooth and has a bounded Hessian)

② $0<\alpha_l\leq\alpha_k\leq\alpha_u<\frac{2}{M}$ for some $\alpha_l, \alpha_u\in\mathbb{R}$.

③ $\min_{x\in\mathbb{R}^n} f(x)+\lambda g(x)$ has a solution.

Then, the sequence $\{x^{(k)}\}_k$ generated by

$$x^{(k+1)} = \text{prox}_{\lambda\alpha_k g}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)$$

satisfies:

(i) $\{f(x^{(k)})+\lambda g(x^{(k)})\}_k$ is monotonically non-increasing,

(ii) $\exists\ u^{(k)}\in\partial g(x^{(k)})$, s.t. $\lim_{k\to+\infty}\|\nabla f(x^{(k)})+u^{(k)}\|_2=0$.

(iii) If $\exists\ B>0$ s.t. $\|x^{(k)}\|_2\leq B\ \forall\ k$, then

$$\lim_{k\to+\infty} f(x^{(k)})+\lambda g(x^{(k)}) = \min_{x\in\mathbb{R}^n} f(x)+\lambda g(x).$$

Remark: The condition $\|x^{(k)}\|_2\leq B$ is satisfied if $f(x)+\lambda g(x)$ is coercive. (Why?).

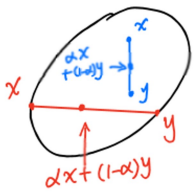
§ 6.3. Constrained Optimization

$$\min_{x \in S} f(x)$$

§ 6.3.1. Convex sets

A set $S \in \mathbb{R}^n$ is convex if

$$x, y \in S \implies \alpha x + (1-\alpha) y \in S \quad \forall\ 0 \le \alpha \le 1$$



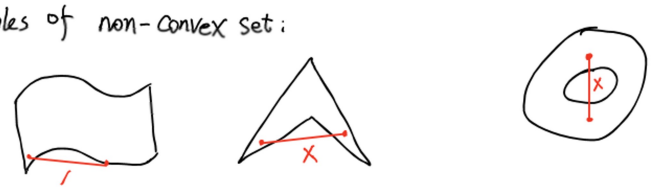
Since $\{\alpha x + (1-\alpha) y \mid 0 \le \alpha \le 1\}$ is the line segment $\overline{xy}$,

a set S is convex if

$\forall\ x, y \in S$, the line segment $\overline{xy}$ in S.

Examples of convex set:

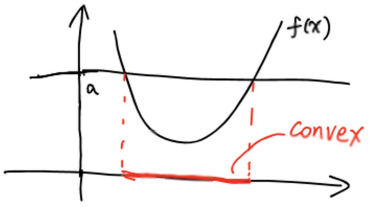
Examples of non-convex set:



More examples of convex and non-convex sets:

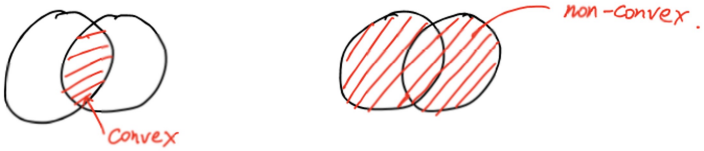
① Solutions of linear equation form a convex set, i.e.,
$S = \{x \in \mathbb{R}^n \mid Ax = b$, where $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^{m \times 1}\}$ is convex for any A and b.

② Level set of a convex function is a convex set, i.e.,
$S = \{x \in \mathbb{R}^n \mid f(x) \le a\}$, where $f: \mathbb{R}^n \to \mathbb{R}$ is convex and $a \in \mathbb{R}$, is convex.



③ Intersection of convex sets is convex.

Union of convex sets is generally not convex.



- If $S \subset \mathbb{R}^n$ is a convex set and $f: \mathbb{R}^n \to \mathbb{R}$ is a convex function, then $\min_{x \in \mathbb{R}^n} f(x)$ s.t. $x \in S$ is a convex optimization.

- Indicator function of a convex set.

Let $S \subset \mathbb{R}^n$ be convex. We define a function $I_S : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

$$I_S(x) = \begin{cases} 0 & \text{if } x \in S \\ +\infty & \text{if } x \notin S \end{cases}$$

This function is called the indicator function of S.

Theorem: If S is a closed convex set, then I_S is a convex function.

proof. Let $x, y \in \mathbb{R}^n$ and $t \in [0,1]$.

Case 1: Both $x, y \in S$.

Then $tx + (1-t)y \in S$ due to the convexity of S.

So, $I_S(tx + (1-t)y) = 0$ and $I_S(x) = I_S(y) = 0$.

Therefore, $I_S(tx + (1-t)y) \leq t\, I_S(x) + (1-t)\, I_S(y)$.

Case 2: $x \notin S$ and $y \in S$.

If $t \neq 0$, $t I_S(x) + (1-t)\, I_S(y) = +\infty \geq I_S(tx + (1-t)y)$

If $t=0$, $tI_S(x)+(1-t)I_S(y) = 0 \geq 0 = I_S(tx+(1-t)y)$

Case 3: $x \in S$, and $y \notin S$.

the same as Case 2

Case 4: $x \notin S$, $y \notin S$.

$t I_S(x) + (1-t) I_S(y) = +\infty \geq I_S(tx+(1-t)y)$

Thus, in all cases, $I_S(tx+(1-t)y) \leq t I_S(x) + (1-t) I_S(y)$ ⊠.

- Sub-differential of indicator function

Let $S \subset \mathbb{R}^n$ be convex. Obviously, I_S is non-differentiable

- If x is in the interior of S, then I_S is differentiable and

$$\partial I_S(x) = \{0\}.$$

- If x is on the boundary of S, then I_S is non-differentiable.

$$\partial I_S(x) = \{ u \in \mathbb{R}^n \mid I_S(y) \geq I_S(x) + \langle u, y-x \rangle \quad \forall y \}$$

Since for any $y \notin S$, $I_S(y) = +\infty$,

$I_S(y) \geq I_S(x) + \langle u, y-x \rangle$ for all $u \in \mathbb{R}^n$.

Therefore, $\partial I_S(x) = \{ u \in \mathbb{R}^n \mid I_S(y) \geq I_S(x) + \langle u, y-x \rangle, \ \forall y \in S \}$

$$\partial I_S(x) = \{ u \in \mathbb{R}^n \mid 0 \geq \langle u, y-x \rangle \quad \forall y \in S \}$$

- If x is not in S, then we define $\partial I_S(x) =$ the empty set.

In summary,

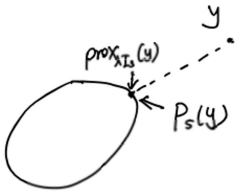
$$\partial I_S(x) = \begin{cases} \text{empty set} & \text{if } x \notin S \\ 0 & \text{if } x \text{ in the interior of } S \\ \{u \in \mathbb{R}^n \mid \langle u, y-x \rangle \leq 0, \ \forall y \in S\} & \text{if } x \text{ on boundary of } S \end{cases}$$

- proximity of the indicator function / projection

By definition,

$$\begin{aligned}\text{prox}_{\lambda I_S}(y) &= \arg\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|x-y\|_2^2 + \lambda I_S(x) \\ &= \arg\min_{x\in S} \tfrac{1}{2}\|x-y\|_2^2 \\ &= \arg\min_{x\in S} \|x-y\|_2\end{aligned}$$

i.e., $\text{prox}_{\lambda I_S}(y)$ is the point on S that is the closest to y, called the projection of y onto S.



From Homework 6, $\text{prox}_{\lambda I_S}(y)$ always exists and is unique.

Furthermore, let $x = \text{prox}_{\lambda I_S}(y)$, and Fermat's lemma gives

$$0 \in x - y + \lambda \partial I_S(x) \iff y - x \in \lambda \partial I_S(x)$$

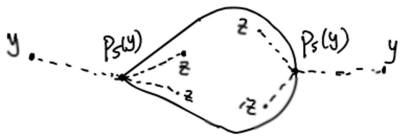
If $y \notin S$, then $y - x \in \lambda \partial I_S(x) = \partial I_S(x)$, i.e.,

$$\langle y-x, z-x\rangle \le 0 \quad \forall z \in S.$$

If $y \in S$, then $x = y$ because $0 \in \lambda \partial I_S(y) \quad \forall y \in S$.

In both cases, $\langle y-x, z-x\rangle \le 0 \quad \forall z \in S$.

Theorem: The projection of y onto S, (i.e., $\text{prox}_{\lambda I_S}(y)$), denoted by $P_S(y)$, always exists and is unique. Furthermore, $P_S(y)$ is the projection if and only if $\langle y - P_S(y), z - P_S(y)\rangle \le 0 \quad \forall z \in S$



Example 1: $S = \{x \mid \|x\|_2 \le 1\}$

S is closed and convex.

- If $\|y\|_2 \le 1$, then $y \in S$ so that $P_S(y) = y$.
- If $\|y\|_2 > 1$, then $y \notin S$, and

$$\forall z \in S,\ \left\langle y - \frac{y}{\|y\|_2},\ z \right\rangle = \left(1 - \frac{1}{\|y\|_2}\right)\langle y, z\rangle \le \left(1 - \frac{1}{\|y\|_2}\right)\|y\|_2\|z\|_2$$

$$\le \left(1 - \frac{1}{\|y\|_2}\right)\|y\|_2 = \left(1 - \frac{1}{\|y\|_2}\right)\left\langle y, \frac{y}{\|y\|_2}\right\rangle = \left\langle y - \frac{y}{\|y\|_2},\ \frac{y}{\|y\|_2}\right\rangle$$

i.e., $\left\langle y - \frac{y}{\|y\|_2},\ z - \frac{y}{\|y\|_2}\right\rangle \le 0$

Therefore, $P_S(y) = \frac{y}{\|y\|_2}$.

Altogether $P_S(y) = \begin{cases} y & \text{if } y \in S \\ \frac{y}{\|y\|_2} & \text{if } y \notin S \end{cases}$

Example 2: $S = \{x \in \mathbb{R}^n \mid \|x\|_\infty \le 1\}$.

S is convex and closed.

Since $S = \{x \in \mathbb{R}^n \mid |x_i| \le 1 \ \ \forall i = 1, 2, \cdots, n\}$, Let $\Omega = \{\alpha \in \mathbb{R} \mid |\alpha| \le 1\}$

$$I_S(x) = \sum_{i=1}^n I_\Omega(x_i)$$

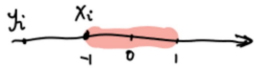
Therefore, $P_S(y) = \arg\min_{x \in S} \|y - x\|_2^2 = \arg\min_{x \in \mathbb{R}^n} \left(\|y - x\|_2^2 + I_S(x)\right)$

$$= \arg\min_{x \in \mathbb{R}^n} \sum_{i=1}^n \left((y_i - x_i)^2 + I_\Omega(x_i)\right)$$

Therefore, $\left[P_S(y)\right]_i = \arg\min_{x_i \in \mathbb{R}} (y_i - x_i)^2 + I_\Omega(x_i)$

$$= \arg\min_{|x_i| \le 1} (y_i - x_i)^2$$

$$= \begin{cases} -1 & \text{if } y_i \le -1 \\ y_i & \text{if } -1 \le y_i \le 1 \\ 1 & \text{if } y_i \ge 1 \end{cases} \equiv P_{[-1,1]}(y_i)$$



Thus, $P_S(y) = \begin{bmatrix} P_{[-1,1]}(y_1) \\ \vdots \\ P_{[-1,1]}(y_n) \end{bmatrix}$

We can check $\langle y - P_S(y),\ z - P_S(y)\rangle \le 0 \quad \forall z \in S$.

§ 6.3.2. Projected gradient descent

Consider $\min_{x \in S} f(x)$, where f is smooth, convex and $S \subset \mathbb{R}^n$ is convex.

This optimization is equivalent to

$$\min_{x \in \mathbb{R}^n} f(x) + I_S(x)$$

We apply Forward-Backward Splitting (FBS) to solve it:

$$x^{(k+1)} = \text{prox}_{\alpha_k I_S}(x^{(k)} - \alpha_k \nabla f(x^{(k)})),$$

i.e.,

$$\boxed{x^{(k+1)} = P_S\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right)}$$

This is known as Projected Gradient Descent. (PGD)

From the theorem in the convergence of FBS,

If $0 < \alpha_l \le \alpha_k \le \alpha_u < \frac{2}{M}$, where $M \in \mathbb{R}$ satisfies $\|\nabla^2 f(x)\|_2 \le M$,

then: ① Since $x^{(k)} \in S$, $\forall k$, $f(x^{(k)}) + I_S(x^{(k)}) = f(x^{(k)})$, $\forall k$,

$\{f(x^{(k)})\}_k$ is monotonically non-increasing.

② If $\|x^{(k)}\|_2 \le B$ $\forall k$ (This is true if, e.g., S is bounded)

then $\lim_{k \to \infty} f(x^{(k)}) = \min_{x \in S} f(x)$.

Example 1: Non-negative least squares

$$\min_{x \ge 0} \|Ax - b\|_2^2$$

Let $S = \{x \in \mathbb{R}^n \mid x \ge 0\}$. Then PGD gives

$$x^{(k+1)} = P_S\left(x^{(k)} - \alpha_k A^T(Ax^{(k)} - b)\right),$$

where P_S is the projection onto S given by

$$P_S(y) = \begin{bmatrix} \max(y_1, 0) \\ \max(y_2, 0) \\ \vdots \\ \max(y_n, 0) \end{bmatrix}.$$

§ 6.4. Case Study: Deep Neural Network Training

Given $\{x^{(i)}, y_i\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

We want to find a function $f: \mathbb{R}^n \to \mathbb{R}$ s.t.

§6.4. Case Study: Deep Neural Network Training

Given $\{x^{(i)}, y_i\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

We want to find a function $f: \mathbb{R}^n \to \mathbb{R}$ s.t.

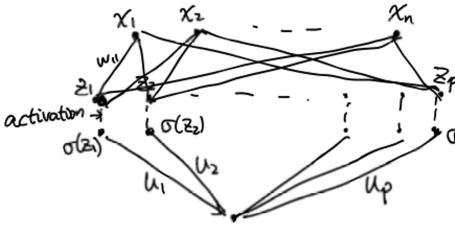
$$f(x^{(i)}) \approx y_i \quad i = 1, 2, \cdots, m.$$

In linear regression, we choose f be an affine function

In kernel regression, we choose f be linear function in the feature space.

In neural network, we choose f be a function generated from a neural network.

- For simplicity, we consider a fully-connected neural network with one hidden layer



Input Layer $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$

Hidden Layer $z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix} \in \mathbb{R}^p$

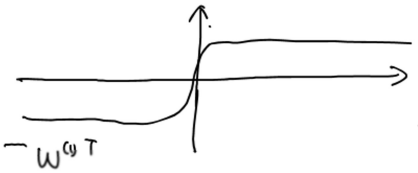
Output layer $y \in \mathbb{R}$

Let $w^{(j)} = \begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jn} \end{pmatrix} \in \mathbb{R}^n$, $u_j \in \mathbb{R}$ $j = 1, 2, \cdots, p$.

Then the input-output can be written as

$$y = \sum_{j=1}^{p} u_j \sigma(\langle w^{(j)}, x \rangle),$$

where $\sigma: \mathbb{R} \to \mathbb{R}$ is an activation function



$$W = \begin{bmatrix} \\ (w^{(p)})^T \end{bmatrix} \in \mathbb{R}^{p\times} , \qquad u = \begin{bmatrix} \\ u_p \end{bmatrix} \in \mathbb{R}$$

Then $y = \langle u, \sigma(Wx)\rangle$, where $\sigma(z) = \begin{bmatrix} \sigma(z_1) \\ \sigma(z_2) \\ \vdots \\ \sigma(z_p) \end{bmatrix}$

So, in neural network, we find W, u,

$$f_{W,u}(x) = \langle u, \sigma(Wx)\rangle$$

s.t.

$$f_{W,u}(x^{(i)}) \approx y_i \qquad i = 1, \cdots, m.$$

Therefore,

> Neural network training
> $\Longleftrightarrow$ Finding W, u s.t. $f_{W,u}(x^{(i)}) \approx y_i$ $i = 1, \cdots, m$.

For this purpose,

$$\min_{W,u} \sum_{i=1}^{m} \left(f_{W,u}(x^{(i)}) - y_i \right)^2$$

Let

$$F_i(W,u) = \left(f_{W,u}(x^{(i)}) - y_i \right)^2$$
$$= \left(\langle u, \sigma(Wx^{(i)})\rangle - y_i \right)^2$$

and

$$F(W,u) = \sum_{i=1}^{m} F_i(W,u)$$

We solve

$$\min_{W,u} F(W,u)$$

Assume σ is differentiable. Then F, F_i are differentiable

A popular algorithm is stochastic gradient descent.

Let [illegible]

for $k = 1, 2, \ldots$

choose an index i randomly

$$(W^{(k+1)}, u^{(k+1)}) = (W^{(k)}, u^{(k)}) - \alpha_k \nabla F_i(W^{(k)}, u^{(k)}),$$

where $\alpha_k > 0$

end

Let us calculate $\nabla F_i(W, u)$, i.e.,

$\nabla_W F_i(W, u)$ and $\nabla_u F_i(W, u)$

- For $\nabla_u F_i(W, u)$:

By the chain rule

$$\nabla_u F_i(W, u) = \nabla_u \Big(\underbrace{(\langle u, \sigma(W x^{(i)}) \rangle}_{f_{W,u}(x^{(i)})} - y_i)^2 \Big)$$

Let $g(t) = t^2$, $G(u) = \langle u, \sigma(W(x^{(i)})) \rangle - y_i$

Then, with W fixed,

$$H(u) \equiv F_i(W, u) = g(G(u))$$

$$\nabla H(u) = \nabla(g(G(u))) = g'(G(u)) \cdot \nabla G(u)$$

$$= 2 G(u) \cdot \sigma(W(x^{(i)}))$$

Therefore,

$$\nabla_u F_i(W, u) = 2 (f_{W,u}(x^{(i)}) - y_i) \cdot \sigma(W x^{(i)})$$

- For $\nabla_W F_i(W, u)$

Let, with u fixed, $H(W) = F_i(W, u)$

Let $g(t) = t^2$, $G(W) = \langle u, \sigma(W x^{(i)}) \rangle$

Then $H(W) = g(G(W) - y_i)$

( )

$$\nabla H(W) = g'(G(W) - y_i)\, \nabla G\, W$$
$$= 2\,(f_{W,u}(x^{(i)}) - y_i) \cdot \nabla G(W)$$

To find $\nabla G(W)$, we find $DG(W)$:

Because $D\,G(W)(V) = D\left(\langle u, \sigma(Wx^{(i)})\rangle\right)(V)$

$$= \langle u,\ D(\sigma(Wx^{(i)}))(V)\rangle$$

To find $D(\sigma(Wx^{(i)}))(V)$, define: $X(W) = Wx^{(i)}$, which is linear in W.

$$D(\sigma \circ X)(W)(V)$$
$$= D\sigma\,(X(W))\Big(DX(W)(V)\Big)$$

$$DX(W)(V) = Vx^{(i)}$$

$$D\sigma(z) = \begin{bmatrix} \sigma'(z_1) & & \\ & \sigma'(z_2) & \\ & & \ddots \\ & & & \sigma'(z_p) \end{bmatrix}$$

$$\sigma(z) = \begin{bmatrix} \sigma(z_1) \\ \sigma(z_2) \\ \vdots \\ \sigma(z_p) \end{bmatrix}$$

$$W = \begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_p^T \end{bmatrix}$$

$$= \begin{bmatrix} \sigma'(w_1^T x^{(i)}) & & \\ & \sigma'(w_2^T x^{(i)}) & \\ & & \ddots \\ & & & \sigma'(w_p^T x^{(i)}) \end{bmatrix} V x^{(i)} \equiv \underline{\mathrm{diag}(\sigma'(Wx^{(i)}))\, V x^{(i)}}$$

Then

$$DG(W)(V) = \langle u,\ \mathrm{diag}(\sigma'(Wx^{(i)}))\, V x^{(i)}\rangle$$
$$= u^T\, \mathrm{diag}(\sigma'(Wx^{(i)}))\, V x^{(i)}$$
$$= \mathrm{trace}\left(u^T\, \mathrm{diag}(\cdot) \cdot V x^{(i)}\right)$$

(($\chi^{(i)}$ 'di ( ) V)

$$= \text{trace} \quad u \quad \text{diag}\cdot$$
$$= \text{trace}\left(\left(\text{diag}(\cdot)\, u(x^{(i)})^T\right)^T V\right)$$
$$= \left\langle \text{diag}\left(\sigma'(Wx^{(i)})\right) u (x^{(i)})^T,\ V \right\rangle$$

Recall the inner product of two matrices

$$\langle W, V \rangle = \text{trace}(W^T V) = \text{trace}(V^T W)$$

For trace $\text{trace}(AB) = \text{trace}(BA)$

$$\Longrightarrow \quad \nabla G(W) = \text{diag}\left(\sigma'(Wx^{(i)})\right) u (x^{(i)})^T$$
$$= \left(\sigma'(Wx^{(i)}) \otimes u\right)(x^{(i)})^T,$$

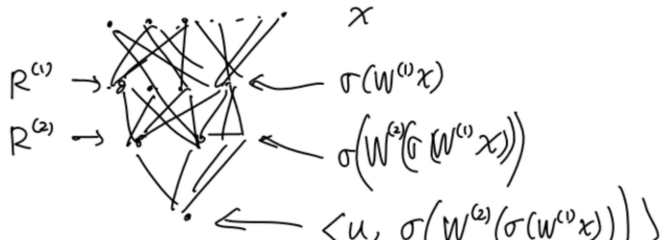
where $\otimes$ means componentwise multiplication

$$\nabla_W F_i(W,u) = 2\left(f_{W,u}(x^{(i)}) - y_i\right) \cdot \left(\sigma'(Wx^{(i)}) \otimes u\right)(x^{(i)})^T$$

Deep Neural Network training

We use L-layer with a large L.

$$f_{W,u}(x) = \langle u, \sigma \circ W^{(L)} \circ \cdots \circ \sigma \circ W^{(2)} \circ \sigma \circ W^{(1)} x \rangle$$



Traing DNN $\Longleftrightarrow$

$$\min_{W,u} \sum_{i=1}^{m} F_i(W,u), \quad \text{where}$$

$$\left[f_{\cdot}(x^{(i)}) - y_i\right]^2$$

$$\left(W \right) \quad \left(\right)$$

$$\int_{\Omega} \left(\frac{\partial w}{\partial t} \right) \, d\Omega$$

$-_i$, $u =$, u .

In order to apply the stochastic gradient descent,
we need to find ∇F_i, i.e.,

$$\nabla_{W^{(l)}} F_i \quad \text{and} \quad \nabla_u F_i$$

$l = 1, \cdots, L$.

For this purpose,
define $R^{(1)}_{W^{(1)}}(x) = \sigma(W^{(1)}x)$

$$R^{(2)}_{W^{(1)},W^{(2)}}(x) = \sigma\left(W^{(2)} R^{(1)}_{W^{(1)}}(x)\right)$$

$$\vdots$$

$$R^{(L)}_{W^{(1)},\cdots,W^{(L)}}(x) = \sigma\left(W^{(L)} R^{(L-1)}_{W^{(1)},\cdots W^{(L-1)}}(x)\right)$$

and $S^{(L)} = \sigma \qquad S^{(L-1)} = \sigma \circ W^{(L)} \circ \sigma, \cdots$

Therefore,

$$f_{w,u}(x) = \langle u, R^{(L)}(x) \rangle$$