Ch.4. Linear functions and Differentiation

§4.1 Linear functions.

Let $f: V \to \mathbb{R}$ be a function on a vector space V.

f is a linear function if

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \forall\, \alpha, \beta \in \mathbb{R} \text{ and } x, y \in V.$$

Example 1: The mean of vectors in $\mathbb{R}^n$. i.e.,

$$f(x) = (x_1 + x_2 + \cdots + x_n)/n, \quad \text{for } x \in \mathbb{R}^n$$

is linear.

Example 2: The maximum element of vectors in $\mathbb{R}^n$, i.e.,

$$f(x) = \max\{x_1, x_2, \cdots, x_n\}, \quad \text{for } x \in \mathbb{R}^n$$

is NOT linear.

Example 3: $f: \mathbb{R}^n \to \mathbb{R}$ with $f(x) = a^T x$ is a linear function.

Example 4: $F: C[a,b] \to \mathbb{R}$ defined by

$$F(f) = f(x), \text{ where } x \in [a,b] \text{ is a given number}$$

is linear.

Example 5: $F: L^2(a,b) \to \mathbb{R}$ defined by

$$F(f) = \int_a^b f(x)\,dx \quad \text{for } f \in L^2(a,b) \text{ is linear.}$$

Example 6: $f: V \to \mathbb{R}$ (V is an inner product space) defined by

$$f(x) = \langle x, z \rangle, \text{ where } z \in V \text{ is a given vector in } V.$$

Example 7: Any norm function on a vector space V is NOT linear.

To see this,

$$\|-x\| = \|x\|,$$

which contradict with

$$f(-x) = f(-x + 0x) = -f(x) + 0f(x) = -f(x)$$

for f being linear on V.

Properties of linear functions

- Homogeneity: $f(\alpha x) = \alpha f(x) \quad \forall\, \alpha \in \mathbb{R}$ and $x \in V$.

  (Because $f(\alpha x) = f(\alpha x + 0x) = \alpha f(x) + 0\, f(x) = \alpha f(x)$)

  It implies $f(0) = 0$, because $f(0) = f(0 \cdot x) = 0 \cdot f(x) = 0 \quad \forall\, x \in V$.

- Additivity: $f(x+y) = f(x) + f(y) \quad \forall\, x, y \in V$.

- $f(\alpha_1 x_1 + \cdots + \alpha_k x_k) = \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k), \quad \forall \alpha_1, \cdots, \alpha_k \in \mathbb{R},\ x_1, \cdots, x_k \in V.$

  To see this, we note that

$$\begin{aligned} f(\alpha_1 x_1 + \cdots + \alpha_k x_k) &= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \cdots + \alpha_k x_k) \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + f(\alpha_3 x_3 + \cdots + \alpha_k x_k) \\ &\ \ \vdots \\ &= \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k). \end{aligned}$$

Inner product representation of a linear function on Hilbert spaces

For simplicity, let's consider a linear function on $\mathbb{R}^n$ equipped with the standard inner product $\langle x, y \rangle = x^T y$ and the induced norm $\|x\|_2 = (\langle x, x \rangle)^{1/2}$.

- From the discussion above,

  For any given $a \in \mathbb{R}^n$, the function $f(x) = \langle a, x \rangle$ is linear.

- The reverse is true, i.e.,

  Any linear function $f : \mathbb{R}^n \to \mathbb{R}$ must be in the form of

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

To see this, let $e_1, e_2, \cdots, e_n$, where $e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow$ ith component, be a basis of $\mathbb{R}^n$,

So that any $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ is written as $x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$.

Therefore, if f is a linear function, then

$$f(x) = f(x_1e_1 + x_2e_2 + \cdots + x_ne_n)$$
$$= x_1 f(e_1) + x_2 f(e_2) + \cdots + x_n f(e_n) \quad \text{— by property of linear functions}$$
$$= \langle a, x \rangle,$$

where $a = \begin{pmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{pmatrix} \in \mathbb{R}^n$.

- Furthermore, the representation of a linear function $f(x) = \langle a, x \rangle$ is unique, which means there is only one vector $a \in \mathbb{R}$ for which $f(x) = \langle a, x \rangle$ holds for all x.

Indeed, suppose that a is not unique, i.e., we have two vectors a, b such that $f(x) = \langle a, x \rangle$ and $f(x) = \langle b, x \rangle$ for all $x \in \mathbb{R}^n$.

Then, let $x = e_i$: $f(e_i) = \langle a, e_i \rangle = a_i$ and $f(e_i) = \langle b, e_i \rangle = b_i$

So, $a_i = b_i$, $i = 1, 2, \cdots, n$.

Therefore $a = b$.

- Altogeter, we see that

> a linear function $f: \mathbb{R}^n \to \mathbb{R}$ if and only if $f(x) = \langle a, x \rangle$ for some unique $a \in \mathbb{R}^n$

✓ The above holds true for any linear function on Hilbert spaces, widely known as Riesz representation theorem.

Theorem (Riesz representation thereom):

> Let H be a Hilbert space, and f be a function: $H \to \mathbb{R}$. Then f is linear and bounded ~~if and only if~~ $f(x) = \langle a, x \rangle$ for some unique $a \in H$.

Example 1: We know mean(x) is linear on $\mathbb{R}^n$. Since $\mathbb{R}^n$ is a Hilbert space, we can find a unique $a \in \mathbb{R}^n$ s.t.

$$\text{mean}(x) = \langle a, x \rangle.$$

Indeed,

$$\text{mean}(x) = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n = \langle a, x \rangle,$$

where $a = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)^T$.

Example 2: Let H be a Hilbert space, and $\|\cdot\|$ is the norm.

It is known that the norm function is NOT linear.

Therefore,

there doesn't exist $a \in H$ such that $\|x\| = \langle a, x\rangle \quad \forall x \in H$.

Hyperplanes

- Again, we consider $\mathbb{R}^n$ as a Hilbert space, where any linear function is written as $\langle a, x\rangle$ for some $a \in \mathbb{R}^n$.

Consider the set

$$S_{a,0} = \{x \in H \mid \langle a, x\rangle = 0\},$$

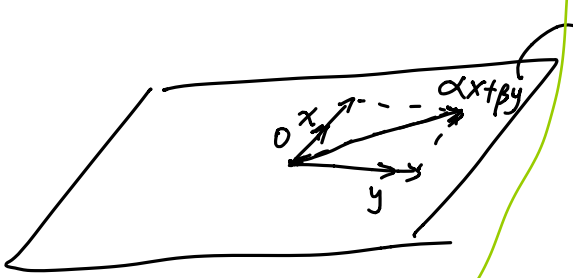
Then, $\forall\, x, y \in S_{a,0}$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle a, \alpha x + \beta y\rangle = \alpha\langle a, x\rangle + \beta\langle a, y\rangle = 0 \Rightarrow \alpha x + \beta y \in S.$$

Therefore, $S_{a,0}$ is a plane.

Since the co-dimension of S_0 is 1 (because it is define by one linear equation)

$S_{a,0}$ is called a hyperplane.



Now let's consider

$$S_{a,b} = \{x \in H \mid \langle a, x\rangle = b\} \quad \text{for } b \in \mathbb{R} \text{ is given.}$$

Let $x_0 \in S_{a,b}$, i.e., $\langle a, x_0\rangle = b$, be fixed.

Then $S_{a,b} = S_{a,0} + x_0$ because:

① $\forall\, x \in S_{a,b}$ $\langle a, x - x_0\rangle = \langle a, x\rangle - \langle a, x_0\rangle = 0$, $\Rightarrow x - x_0 \in S_{a,0}$.

$\Rightarrow x \in S_{a,0} + x_0$. $\Rightarrow$ $S_{a,b} \subseteq S_{a,0} + x_0$ ①

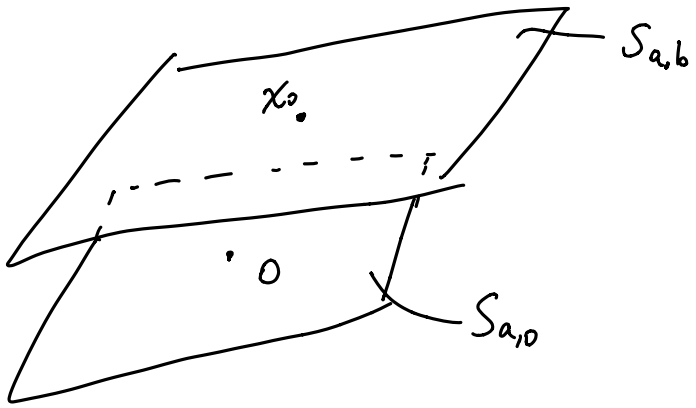
② $\forall x \in S_{a,0}$ $\langle a, x+x_0\rangle = \langle a,x\rangle + \langle a, x_0\rangle = b \Rightarrow x+x_0 \in S_{a,b}$

$S_{a,0} + x_0 \subseteq S_{a,b}$ ②

In other words,

$S_{a,b}$ is a shift of a hyperplane, still called a hyperplane.

① ② $\Rightarrow S_{a,0}+x_0 = S_{a,b}$



- This concept can be generalized to any inner product space V.

The set $\{x \in V \mid \langle a,x\rangle = b\}$, where $a \in V$ and $b \in \mathbb{R}$ are given, is called a Hyperplane in V.

Projection onto hyperplanes

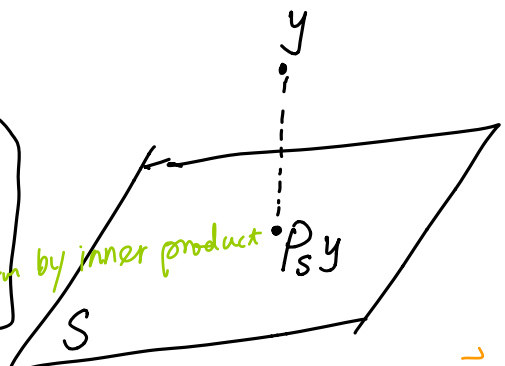
- Consider a Hilbert space V and a hyperplane S

$$S = \{x \in V \mid \langle a,x\rangle = b\}$$

Let $y \in V$ be a given vector.

The vector on S that is the closest to y is called the projection of y on S, denoted by $P_S y$,

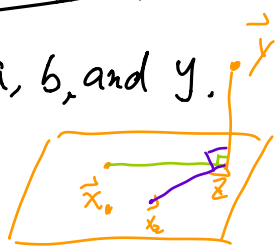
i.e., $P_S y = \arg\min_{x \in S} \|x-y\|$ → induced norm by inner product



- Let us find an explicit expression of $P_S y$ in terms of a, b, and y.

Theorem: z is a solution of $\min_{x \in S} \|x-y\|$ if and only if

$z \in S$ and $\langle z-y, z-x\rangle = 0$ $\forall x \in S$.



Proof. ① We first prove that: If $z \in S$ is a solution of $\min_{x \in S} \|x-y\|$,

then $\langle z-y, z-x\rangle = 0$ $\forall x \in S$.

Since z is a solution, $z \in S$, i.e., $\langle a, z\rangle = b$.

$\forall x \in S$ and $t \in \mathbb{R}$, it is easy to see that

$$\langle a,\ (1+t)z - tx\rangle = (1+t)\langle a, z\rangle - t\langle a, x\rangle = b.$$

Therefore, $(1+t)z - tx \in S$. (构造一个仍在 $S_{a,b}$ 中含有 z 和 x 之间的点 $(1+t)\vec{z} - t\vec{x}$)

Since z is closest to y on S, we have

$$\|z-y\|^2 \le \|[(1+t)z - tx] - y\|^2$$
$$= \|(z-y) + t(z-x)\|^2$$
$$= \|z-y\|^2 + t^2\|z-x\|^2 + 2t\langle z-y,\ z-x\rangle.$$

i.e., $$t\langle z-y,\ z-x\rangle \ge -\frac{t^2}{2}\|z-x\|^2. \quad (*)$$

i.e. 现在我们证明(*)式对 $t \in \mathbb{R}$ 都成立

现(*)至少得在 $t \to 0^+$ 和 $t \to 0^-$ 时成立：即：① ②

- If we choose $t > 0$,

$$\langle z-y,\ z-x\rangle \ge -\frac{t}{2}\|z-x\|^2$$

① Letting $t \to 0_+$, then $\langle z-y,\ z-x\rangle \ge 0$.

- If we choose $t < 0$,

$$\langle z-y,\ z-x\rangle \le -\frac{t}{2}\|z-x\|^2$$

② Letting $t \to 0_-$, then $\langle z-y,\ z-x\rangle \le 0$

Altogether, z satisfies $\langle z-y,\ z-x\rangle = 0$ must be $\forall x \in S$.

② We then show that $z \in S$ satisfies $\langle z-y, z-x\rangle = 0$, then z is a solution of $\min_{x\in S}\|x-y\|$, by direct calculation.

Since $\langle z-y, z-x\rangle = 0 \quad \forall x \in S$,

$$\|x-y\|^2 = \|(z-x)-(z-y)\|^2$$
$$= \|z-x\|^2 + \|z-y\|^2 - 2\langle z-x, z-y\rangle$$
$$= \|z-y\|^2 + \|z-x\|^2 \ge \|z-y\|^2 \quad \forall x \in S.$$

This, together with $z \in S$, implies z minimizes $\|x-y\|^2$ in $x \in S$.

H
↓
$\langle\cdot,\cdot\rangle$
↓
$\|\ \| = \sqrt{\langle x,x\rangle}$ i.e. $\|x\|^2 = \langle x,x\rangle$

在H中，$\sqrt{\langle x,x\rangle} \Rightarrow \|\ \|$

所以看到 $\|x\|^2$ 想成 $\langle x,x\rangle$

Theorem: The solution of $\min_{x\in S}\|x-y\|^2$ exists and unique, which is given by

$$y-\left(\frac{\langle a,y\rangle-b}{\|a\|^2}\right)a$$

proof. denote $z=y-\left(\frac{\langle a,y\rangle-b}{\|a\|^2}\right)a$.

① $\langle a,z\rangle = \langle a,y\rangle - \left(\frac{\langle a,y\rangle-b}{\|a\|^2}\right)\langle a,a\rangle$

$= \langle a,y\rangle - (\langle a,y\rangle - b) = b$, so $z\in S_{a,b}$

② $\forall\, x\in S$,

$$\langle z-y, z-x\rangle = -\frac{\langle a,y\rangle-b}{\|a\|^2}\langle a, z-x\rangle$$

$$= -\frac{\langle a,y\rangle-b}{\|a\|^2}\left(\langle a,z\rangle - \langle a,x\rangle\right) = 0$$

(because $\langle a,z\rangle = \langle a,x\rangle = b$).

By the previous theorem, z is a solution of $\min_{x\in S}\|x-y\|$.

It remains to show the uniqueness.

Suppose we have two solutions z_1 and z_2. Then,

z_1 is a solution, $\Rightarrow \langle z_1-y, z_1-z_2\rangle = 0$ ①

z_2 is a solution, $\Rightarrow \langle z_2-y, z_2-z_1\rangle = 0$ ②

Taking difference leads to $\langle z_1-z_2, z_1-z_2\rangle = 0$

$\Rightarrow \|z_1-z_2\|^2 = 0 \Rightarrow z_1 = z_2$ ⊗

- In summary, the projection $P_S y$ of $y\in V$ onto the hyperplane

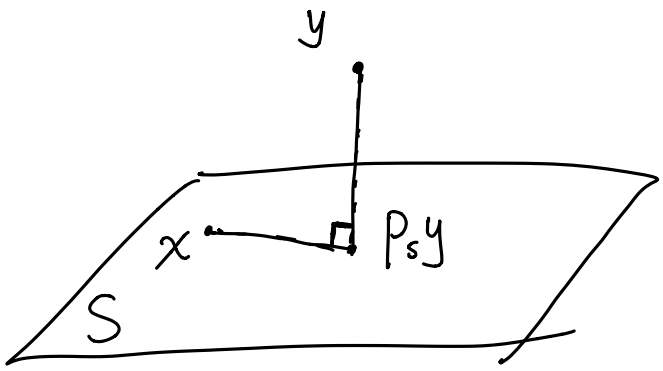
$$S=\{x\in V \mid \langle a,x\rangle = b\}$$

exists and is unique. Furthermore,

$$P_S y = y-\left(\frac{\langle a,y\rangle-b}{\|a\|^2}\right)a$$

and it satisfies

$$\langle P_S y - y, P_S y - x\rangle = 0.$$



Affine functions

A linear function plus a constant is called an affine function.

That is, a function $f: V \to \mathbb{R}$ is affine if

$$f(x) = g(x) + b,$$

where $g: V \to \mathbb{R}$ is linear and $b \in \mathbb{R}$ is a constant.

Properties:

- If $f: V \to \mathbb{R}$ is affine, then

  $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ $\forall\, x, y \in V$ and $\alpha, \beta \in \mathbb{R}$ s.t. $\alpha + \beta = 1$.

  To see this, (g linear, $\alpha+\beta=1$)

  $$f(\alpha x + \beta y) = g(\alpha x + \beta y) + b = \alpha g(x) + \beta g(y) + (\alpha+\beta) b$$
  $$= \alpha(g(x)+b) + \beta(g(y)+b) = \alpha f(x) + \beta f(y).$$

- If $f: \mathcal{H} \to \mathbb{R}$ (f is affine), where $\mathcal{H}$ is a Hilbert space, then

  f must be in the form of

  $$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in V \text{ and } b \in \mathbb{R}.$$

§4.2 Case Studies: Regression and Classification.

§4.2.1 Linear Regression:

- Given a set of data

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_N, y_N),$$

where

$x_i \in \mathbb{R}^n$ is an input feature vector
$y_i \in \mathbb{R}$ is the corresponding response to x_i.
, $i = 1, 2, \cdots, N$.

Given a new input feature vector $x \in \mathbb{R}^n$, how to predict the corresponding response $y \in \mathbb{R}$?

For example, $x_i \in \mathbb{R}^n$ represents n attributes of a house, and $y_i \in \mathbb{R}$ is the selling price.

We want to predict the selling price of a house with feature $x \in \mathbb{R}^n$.

- Mathematically, we need to find a function $f: \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x_i) \simeq y_i \quad , \quad i = 1, 2, \cdots, N$$

This is called regression. In this context,

x_i, are called regressor / independent varibles.
y_i are called dependent variables / outcome / label.

- The class of all functions $\mathbb{R}^n \to \mathbb{R}$ is too large (包含太多我们不关心的), and the given data set $\{(x_i, y_i)\}_{i=1}^N$ is not enough to determine a function uniquely.

So, we need to find a function class Φ where we search f.

~~Intuitively, larger N, larger function class Φ.~~

- Linear ~~model~~ regression: We choose as Φ.

We search f in the class of all affine functions,

i.e., $f(x) = \langle a, x \rangle + b$ for some $a \in \mathbb{R}^n$, $b \in \mathbb{R}$.

(in this problem, $\mathbb{R}^n$ is our H, so, $f(x) \in \Phi$ can be denoted as)

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

$$\langle a, x_i \rangle + b \approx y_i, \quad i = 1, 2, \cdots, N,$$

by minimizing the error of the linear equations.

While there are many possible definitions of error, it is popular to consider the square error as follows:

$$\left(\langle a, x_i \rangle + b - y_i\right)^2, \quad i = 1, 2, \cdots, N.$$

Therefore, we find $a \in \mathbb{R}^n, b \in \mathbb{R}$ by solving

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

This problem is called the Least squares (LS) problem.

Write $X = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_N^T & 1 \end{bmatrix} \in \mathbb{R}^{N \times (n+1)}$, $\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$

and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$.

Then LS problem becomes

$$(*) \quad \boxed{\min_{\beta \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2}$$

Since we have N linear equations to fit and $n+1$ unknowns,

$N \geq n+1$ is required [to have a unique solution].

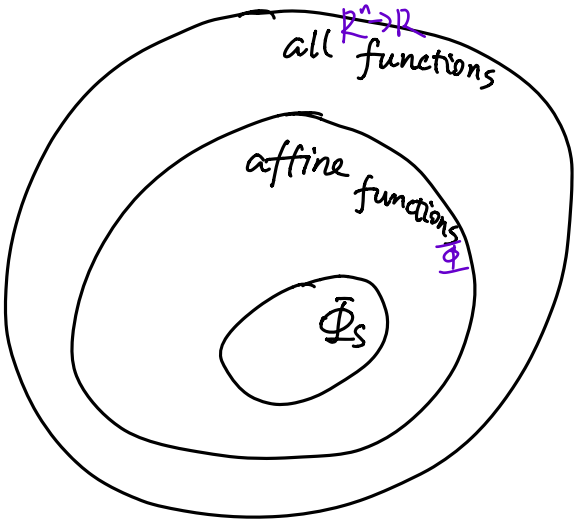
However, in practice, N can be much much smaller than n.

For example, x_i are images of size 10M pixels (i.e., $n = 10M$), and it is very difficult to have a data base of 10M images (i.e., $N << 10M$).

- Regularization.

  In many applications, we don't have enough data (e.g., $N < n+1$), so that the class of affine functions is too large to search f.

  We search f in a sub class of all affine functions



Therefore, instead of searching $\beta \in \mathbb{R}^{n+1}$, we search $\beta \in S \subset \mathbb{R}^{n+1}$

$$\min_{\beta \in S} \|X\beta - y\|_2^2$$

In other words, we search f in

$$\Phi_S = \{ f: \mathbb{R}^n \to \mathbb{R} \mid f(x) = \langle a, x \rangle + b, \ \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in S \}$$

By choosing different S, we obtain different approaches.

— Ridge regression:

We choose $S = \{ \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_2 \le C \}$ for some $C > 0$.

Then we solve $\min_{\beta \in S} \|X\beta - y\|_2^2$

$\Leftrightarrow$ by convex optimization theory

$$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda \|a\|_2^2,$$

where $\lambda > 0$ depends on C and others.

Here $\|a\|_2^2$ is the regularization term.

and $\frac{1}{2}\|X\beta - y\|_2^2$ is the data-fitting term.

In other words, we find $\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$ such that the error of data fitting and the $\|a\|_2^2$ are minimized simultaneously.

larger λ, smaller Φ_S, looser fitting to the data

smaller λ, larger Φ_S, tighter fitting to the data

Therefore, ridge regression gives $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

$X\beta \approx y$ and $\|a\|_2$ is small. (so that $\beta \in S$)

The parameter λ is tune s.t. $\|a\|_2 \le C$.

— LASSO regression.

We choose $S = \{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_1 \le C\}$

The we solve $\min_{\beta \in S} \|X\beta - y\|_2^2$

$\Updownarrow$ ← by convex optimization theory

$$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda \|a\|_1$$

where $\lambda > 0$ depends on C and others.

Here the regularization term is $\|a\|_1$.

So, we find $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

the error of data fitting and the $\|a\|_1$

are minimized simultaneously.

✓ Small $\|a\|_1$ tends to give a sparse vector $a \in \mathbb{R}^n$

(i.e., many entries of a are zeros)

✓ Consequently, LASSO gives a solution $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

$X\beta \approx y$ and a is sparse

Thus, given $x \in \mathbb{R}^n$, the prediction is given by

$$\langle x, a \rangle + b = \sum_{i=1}^{n} a_i x_i + b = \sum_{i \in I} a_i x_i + b$$

Let $I = \{i \mid a_i \neq 0\}$

Since I is a small set, the prediction depends only on

a small portion entries of x.

This is preferred because the prediction is interpretable.

可解釋

§4.2.2. Kernel regression

Again, linear regression has its limitation.

We extend it to nonlinear regression by Kernel trick.

Feature map $\phi: \mathbb{R}^n \to H$ (H is some Hilbert space)

Then do regression in H.

However, since H is very large, the set of all linear functions is also too large. We need regularization.

The task is, Find a linear and bounded function $f: H \to \mathbb{R}$, s.t. $f(\phi(\vec{x}_i)) \simeq y_i$ (no need for affine)

We solve

$$\min_{a \in H} \frac{1}{2} \sum_{i=1}^{N} \left(\langle a, \phi(x_i) \rangle_H - y_i \right)^2 + \lambda \|a\|_H^2$$

($\langle \cdot, \cdot \rangle_H$: inner prod in H; $\|a\|_H^2$: induced norm by $\langle , \rangle$; $\lambda > 0$ is a ...; as f is lin and bounded)

(KR)

Representer Theorem:

The solution must be in the form of $\vec{a} = \sum_{i=1}^{N} c_i \phi(x_i)$ for some $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

($\langle \vec{a}, \phi(\vec{x}_i) \rangle_H$, $(\vec{a} \in H)$)

✓ H is usually infinite.

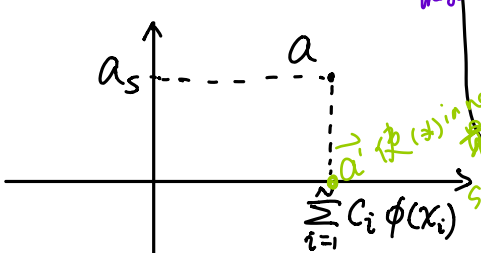
✓ N data

Proof. For any $a \in H$, we claim that a can be decomposed as

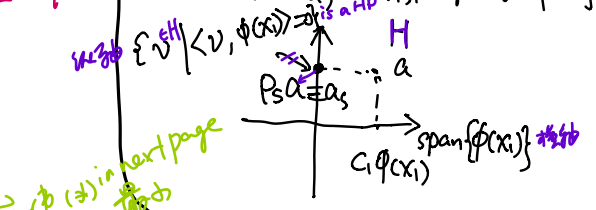
$$a = a_S + \sum_{i=1}^{N} c_i \phi(x_i)$$

where $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$ and $\langle a_S, \phi(x_i) \rangle = 0$ for $i = 1, 2, \cdots, N$.

(c: what we want to determine)



Indeed consider $N=1$ for simplicity.



Then, $S = \{v \mid \langle v, \phi(x_i) \rangle = 0\}$ is a Hyperplane. (with $\vec{0}$ in it)

For any $a \in H$

$$a = P_S a + (a - P_S a),$$

where $P_S a$ is the projection of a onto S.

So, by the property of projection,

$$\langle P_S a - \vec{a}, P_S a - \vec{0} \rangle = 0 \quad (\text{because } \vec{0} \in S)$$

$$\Rightarrow \langle P_S a - a, P_S a \rangle = 0$$

Also, by direct calculation,

$$a - P_S a = c_1 \phi(x_1)$$

So, $a = P_S a + c_1 \phi(x_1)$

For general N, it can be done similarly.

$$P_S y = y - \left(\frac{\langle a, y \rangle - b}{\|a\|^2} \right) a$$

$$P_S a = a - \left(\frac{\langle \phi(x_1), a \rangle - 0}{\|\phi(x_1)\|^2} \right) \phi(x_1)$$

Therefore,

$$\frac{1}{2}\sum_{i=1}^{N}\left(\langle a, \phi(x_i)\rangle - y_i\right)^2 + \lambda \|a\|_H^2 \quad (*)$$

(a can be decomposed to $a_S + \sum_{i=1}^{N} c_i\phi(x_i)$)

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\langle \sum_{j=1}^{N} c_j\phi(x_j) + a_S, \phi(x_i)\rangle - y_i\right)^2 + \lambda \left\| \sum_{i=1}^{N} c_i\phi(x_i) + a_S\right\|_H^2$$

(② $a_S \perp \phi(x_i), \forall i \in 1{:}N$)

① $\langle\cdot,\cdot\rangle$ is linear

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j\langle\phi(x_j), \phi(x_i)\rangle - y_i\right)^2 +$$

$$\lambda\left(\langle \sum_{i=1}^{N} c_i\phi(x_i), \sum_{j=1}^{N} c_j\phi(x_j)\rangle + 2\langle a_S, \sum_{i=1}^{N} c_i\phi(x_i)\rangle + \langle a_S, a_S\rangle\right)$$

($\|\star\|_H^2 = \langle\star,\star\rangle_H$; $\langle a_S, \sum c_i\phi(x_i)\rangle = 0$)

let $\langle\phi(x_i),\phi(x_j)\rangle = K(x_i,x_j)$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j K(x_i,x_j) - y_i\right)^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i,x_j) + \lambda\|a_S\|_H^2$$

(best approximation)

$$= \frac{1}{2}\|K\vec{c} - \vec{y}\|_2^2 + \lambda c^T K c + \lambda\|a_S\|_H^2$$

($\|\cdot\|_2$: $\mathbb{R}^N$ 2-norm)

where $K = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_N) \\ \vdots & & & \\ K(x_N,x_1) & K(x_N,x_2) & \cdots & K(x_N,x_N) \end{bmatrix} \in \mathbb{R}^{N\times N}$ $\quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

Let $F_1(c) = \frac{1}{2}\|Kc - y\|_2^2 + \lambda c^T K c$ —— depends on $c \in \mathbb{R}^N$ only.

$F_2(a_S) = \lambda\|a_S\|_H^2$ —— depends on $a_S \in H$ only.

Then, the minimization is the same as

$$\min_{\substack{c\in\mathbb{R}^N \\ a_S\in H \\ \langle a_S,\phi(x_i)\rangle=0,\ i=1,\cdots,N}} F_1(c) + F_2(a_S)$$

$$\Updownarrow \quad a = a_S + \sum_{i=1}^{N} c_i\phi(x_i)$$

$$\min_{c\in\mathbb{R}^N} F_1(c) \quad \text{and} \quad \min_{\substack{a_S\in H \\ a_S\text{满足}\langle a_S,\phi(x_i)\rangle=0 \\ i=1,\cdots,N}} F_2(a_S)$$

(显然满足上于 $\phi(x_i)$)

Obviously, because $\|a_S\|_H^2 \geq 0$, $\min_{\substack{a_S\in H \\ \langle a_S,\phi(x_i)\rangle=0}} F_2(a_S)$ is solved by $a_S = \vec{0}$.

Thus, the solution of the original minimization is

$$a = \sum_{i=1}^{N} c_i\phi(x_i) \quad (KR)$$

where $c \in \mathbb{R}^N$ is a solution of $\min_{c\in\mathbb{R}^N} F_1(c)$. ⊠

From the above proof, we see that

$$\min_{a \in H} \frac{1}{2} \sum_{i=1}^{N} \left(\langle a, \phi(x_i) \rangle - y_i\right)^2 + \lambda \|a\|_H^2$$

$$\Updownarrow$$

$$\boxed{\min_{c \in \mathbb{R}^N} \frac{1}{2} \|Kc - y\|_2^2 + \lambda c^T K c}$$

Let the solution be $c \in \mathbb{R}^N$.

Then the predicted output y for input $x \in \mathbb{R}^n$ is

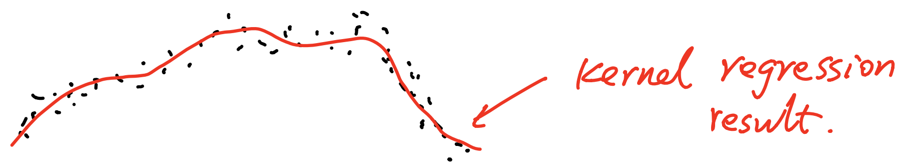
$$y = \langle a, \phi(x) \rangle = \langle \sum_{i=1}^{N} c_i \phi(x_i), \phi(x) \rangle$$

$$= \sum_{i=1}^{N} c_i K(x_i, x).$$

$K: (\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$

All the computation involves only the kernel function $K(\cdot, \cdot)$,

So, No explicit feature map $\phi(\cdot)$ is needed.



整理下 kernel (Ridge) regression Alg: @ 10-17. 1.02'31

① Choose a kernel function.

$$K(\vec{x}, \vec{y}) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

e.g. $K(\vec{x}, \vec{y}) = e^{-\frac{\|\vec{x} - \vec{y}\|_2^2}{\delta^2}}$ (Gaussian Kernel)

$K(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + 1)^d$ (polynomial kernel)

② Calculate the Kernel matrix

$$K = \left[K(\vec{x}_i, \vec{x}_j)\right]_{i,j=1}^{N}, \in \mathbb{R}^{N \times N}$$

→ no. of samples

③ Solve $\vec{c}^*$ from

$$\vec{c}^* = \arg\min_{\vec{c} \in \mathbb{R}^N} \left(\|K^T \vec{c} - \vec{y}\|_2^2 + \lambda \vec{c}^T K \vec{c}\right)$$

④ Then the regression function is

$$f(\vec{x}) = \langle \vec{a}^*, \phi(\vec{x}) \rangle_H = \langle \sum_{i=1}^{N} c_i^* \phi(\vec{x}_i), \phi(\vec{x}) \rangle_H = \sum_{i=1}^{N} c_i^* K(\vec{x}_i, \vec{x})$$ ☒

这里也解释了我们为什么要把 a^* 写成这样的形式：以构造出 $K(\cdot,\cdot)$

§4.2.3 Linear Classification

- Classification: Giving training data

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N), \quad x_i \in \mathbb{R}^n, \quad y_i \in \{-1, +1\}, \quad i = 1, \cdots, N,$$

find a classifier (a function) f such that

$$y_i = \begin{cases} 1, & \text{if } f(x_i) \geq 1 \\ -1, & \text{if } f(x_i) \leq -1 \end{cases}.$$

We use hyperplanes to separate the points

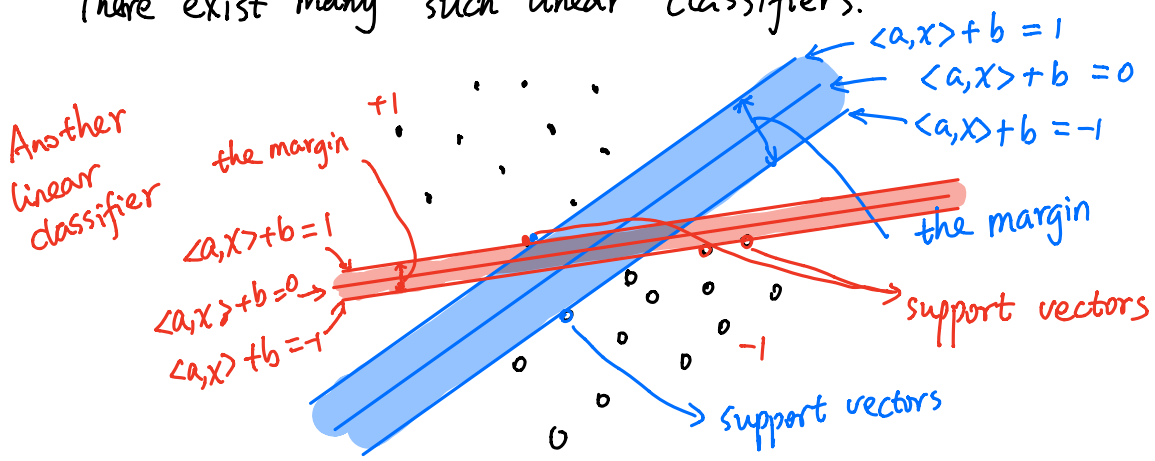
$$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in \mathbb{R}^n, \quad b \in \mathbb{R}.$$

The weights $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are noralized such that

$$\langle a, x_i \rangle + b_i \begin{cases} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{cases}$$

- Support Vector Machine (SVM)

There exist many such linear classifiers.



Which one is the better?

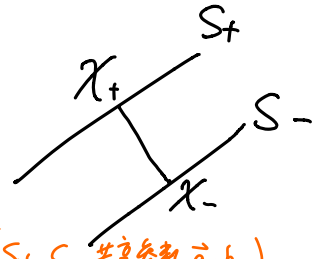
From the figure, the blue is better, because it has a larger margin, and hence a larger buffer zone of mis-classification.

Therefore, we want to maximize the margin among all candidates.

Let us calculate the margin in terms of a and b.

The margin is the distance between the two hyperplanes

$$S_+ = \{x \mid \langle a, x \rangle + b = 1\} \text{ and } S_- = \{x \mid \langle a, x \rangle + b = -1\}$$



(S_+, S_- 共享參數 a, b)

Let $x_+ \in S_+$ and $x_- \in S_-$ such that

$$\|x_+ - x_-\|_2 = \text{dist}(S_+, S_-).$$

Since x_+ is a projection of x_- onto $S_+ = \{x \mid \langle a, x\rangle = 1-b\}$ ($\langle a, x\rangle + (b-1) = 0$)

$$x_+ = x_- - \frac{\langle a, x_-\rangle + b - 1}{\|a\|_2^2}\vec{a}$$

$$= x_- - \frac{-1-b+b-1}{\|a\|^2}a \quad (\text{since } \vec{x}_- \in S_-)$$

$$= x_- + \frac{2}{\|a\|_2^2}a$$

Thus, $\|x_+ - x_-\|_2 = \left\|\frac{2}{\|a\|_2^2}a\right\|_2 = \frac{2}{\|a\|_2}$

i.e., the margin $= \frac{2}{\|a\|_2}$

2-norm, $\left\|\frac{2}{\|\vec{a}\|_2^2}\vec{a}\right\|_2 = \frac{2}{\|\vec{a}\|_2^2}\|\vec{a}\|_2 = \frac{2}{\|\vec{a}\|_2}$

Support Vector Machine (SVM)

$$\max_{a\in\mathbb{R}^n,\, b\in\mathbb{R}} \ \frac{2}{\|a\|_2}$$

$$\text{s.t.} \quad \langle a, x_i\rangle + b \geq 1 \ \text{ if } \ y_i = 1$$

$$\langle a, x_i\rangle + b \leq -1 \ \text{ if } \ y_i = -1 \ ,$$

which is equivalent to

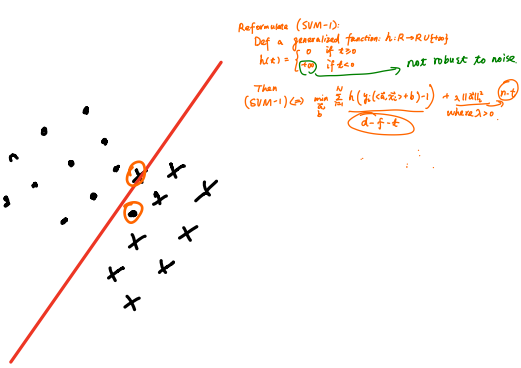
$$\min_{a\in\mathbb{R}^n,\, b\in\mathbb{R}} \ \frac{1}{2}\|a\|_2^2 \qquad \text{(SVM-1)}$$

$$\text{s.t} \quad y_i(\langle a, x_i\rangle + b) \geq 1$$

- The above SVM is NOT robust to noise.

For example shown on the right,

even we have only two noisy points, there is no solution to (SVM-1).



We consider a "soft" version of (SVM-1) as follows.

We minimize the error to the separation

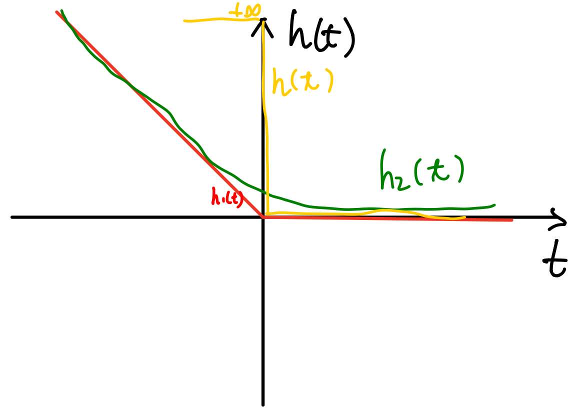
$$\sum_{i=1}^{N} h\left(\left(y_i(\langle a, x_i\rangle + b)\right) - 1\right),$$

where the function

$$h(t) = \begin{cases} 0 & \text{if } t \geq 0 \\ |t| & \text{if } t < 0 \end{cases} \qquad = \max(0, -t)$$

not $+\infty$ more robust

$h_2(t) = \ln(e^0 + e^{-t})$ (soft max), smoother approximate (max)



In other words, if $y_i\langle a, x_i\rangle + b \geq 1$, the error is 0. *no longer the +∞.*

if $y_i\langle a, x_i\rangle + b < 1$, the error is *the absolute value of $y_i\langle a, x_i\rangle + b_i - 1$.*

Also, the margin $\frac{1}{2}\|a\|_2^2$ can be viewed as a regularization.

Altogether, we solve

$$\min_{\substack{a\in\mathbb{R}^n \\ b\in\mathbb{R}}} \sum_{i=1}^{N} h\left(y_i(\langle a, x_i\rangle + b_i) - 1\right) + \tfrac{1}{2}\|a\|_2^2 \qquad (\text{SVM-2})$$

($\frac{1}{2}$ → $\lambda > 0$)

We call this SVM with a soft margin. *Soft-Margin SVM*

As (SVM-2) is not smooth (10.14, 2'30)

h can be approximated by some smooth function. If h is chosen the so-called logistic function, we call it logistic regression. *($h_2(t)$ 更好.)*

- Kernel SVM

The linear SVMs don't work for curved data.

The kernel method can be used.

Let $\phi : \mathbb{R}^n \to H$ be a feature map.

So, we use linear functions on H to classify the points

By Riesz representation theorem, any linear function in the form of

$\langle a, x\rangle$ for some $a \in H$.

和上面一样 在H中不再需要使用affine function 只需linear function and bounded

Therefore, (SVM-2) becomes

$$\min_{a\in H} \tilde{h}\left(y_i\langle a, \phi(x_i)\rangle - 1\right) + \tfrac{1}{2}\|a\|_H^2 \qquad (\text{K-SVM})$$

($\frac{1}{2}$ → $\lambda > 0$)

$\tilde{h}$ can be chosen as h or h_1 or h_2

*→ 于是a是唯一待定参数

Aagain, one can prove the following representer theorem.

Theorem: Any solution of (K-SVM) is in the form of

$$a = \sum_{i=1}^{N} c_i \phi(x_i)$$ (样本在H中投影的线性组合)

proof. ① Write $a = \sum_{i=1}^{N} c_i \phi(x_i) + a_S$ for some $a_S \in H$ and $\langle a_S, \phi(x_i)\rangle = 0 \ \forall i$. (P13)

② The rest is the almost same as the linear regression case. ⊠ (P14)

Thus, 把 a 代入 (K-SVM), we have becomes

$$\min_{c \in \mathbb{R}^N} \tilde{h}\left(y_i \left(\sum_{j=1}^{N} c_j K(x_i, x_j) c_j \right) - 1 \right) + \frac{1}{2} c^T K c,$$

where $K = [K(x_i, x_j)]_{i=1,j=1}^{N,N} \in \mathbb{R}^{N \times N}$.

The prediction of the input x is given by $\operatorname{sgn}(\langle a, \phi(x_i)\rangle_H)$

$$\operatorname{sgn}\left(\sum_{j=1}^{N} K(x, x_j) c_j \right) \Leftarrow \text{代入 } \vec{a} = \operatorname{sgn}\left(\left\langle \sum_{j=1}^{N} c_j \phi(x_j), \phi(x_i) \right\rangle \right)$$

Again, only $K(\cdot,\cdot)$ is needed in the kernel SVM, and no explicit feature maps $\phi(\cdot)$ is required.

Full alg:

① Def. a $K: (\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$ ($\vec{x}_i \in \mathbb{R}^n$, $i = 1, \cdots, N$)

② Calculate K matrix with samples $(\vec{x}_i)$:

$$K = [K(x_i, x_j)]_{i,j=1}^{N} \in \mathbb{R}^{N \times N}$$

③ Solve

$$c^* = \operatorname*{argmin}_{\vec{c} \in \mathbb{R}^N} \sum_{i=1}^{N} \tilde{h}\left(y_i (K\vec{c})_i - 1 \right) + \lambda c^T K c \quad (+\lambda_1 \|\vec{a}_S\|_H^2)$$

($(K\vec{c})_i$: i-th entry; $\|\vec{a}_S\|_H^2 \to \vec{0}$)

④ $a^* = \sum_{j=1}^{N} c_j^* \phi(x_j)$

§4.3. Linear approximation and Differentiation.

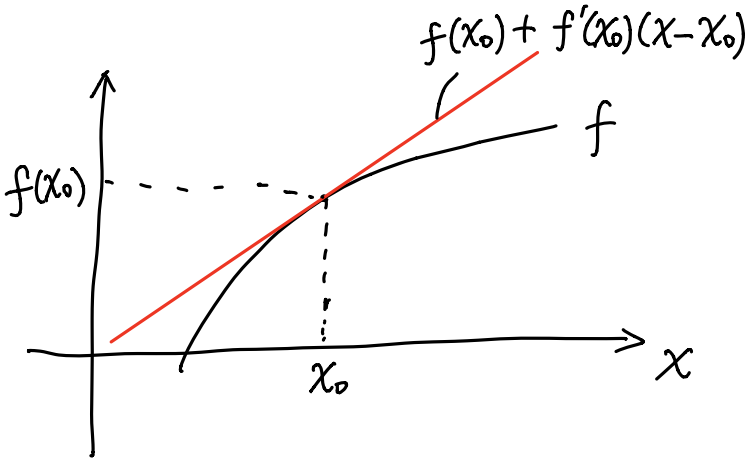
Recall that for a function $f: \mathbb{R} \to \mathbb{R}$, the derivative at x_0 is

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which is the same as

$$\lim_{x \to x_0} \left| \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} \right| = 0$$

Notice that $f(x_0) + f'(x_0)(x - x_0)$ is an affine function in $\mathbb{R}$ that passes through $(x_0, f(x_0))$.



In other words, in differentiation at x_0,

(1) f is approximated by an affine function passes thru $(x_0, f(x_0))$.

(2) the error of the approximation is $o(|x - x_0|)$.

(o : little o, i.e., $\lim_{x \to x_0} \frac{|\text{error}|}{|x - x_0|} \to 0$)

This can be used to define differentiation of functions on Hilbert spaces. (We use Hilbert space for simplicity. It can be easily adapted to Banach spaces.)

Let $f: V \to \mathbb{R}$ with V a Hilbert space.

Consider the differentiation of f at $x^{(0)} \in V$.

(1) By Riesz representation theorem, any affine function is in the form

of $\langle v, x\rangle + a$ for some $v \in V$ and $a \in \mathbb{R}$. Since it passes thru $(x^{(0)}, f(x^{(0)}))$, $\langle v, x^{(0)}\rangle + a = f(x^{(0)})$. Therefore, the affine function is in the form of

$$\langle v, x\rangle + a = \langle v, x - x^{(0)}\rangle + (\langle v, x^{(0)}\rangle + a)$$
$$= f(x^{(0)}) + \langle v, x - x^{(0)}\rangle.$$

(2). The approxition error is

$$\text{error} = \left| f(x) - f(x^{(0)}) - \langle v, x - x^{(0)}\rangle \right|,$$

The error should be in the order of $o(\|x - x^{(0)}\|)$, i.e.,

$$\frac{\text{error}}{\|x - x^{(0)}\|} \to 0 \quad \text{as} \quad x \to x^{(0)} \ (\text{i.e., } \|x - x^{(0)}\| \to 0)$$

Definition: Let V be a Hilbert space. Let $f: V \to \mathbb{R}$. Then f is said Frechet differentiable if there exists a $v \in V$ such that

$$\lim_{x \to x^{(0)}} \frac{\left| f(x) - f(x^{(0)}) - \langle v, x - x^{(0)}\rangle \right|}{\|x - x^{(0)}\|} = 0. \quad \left(\begin{array}{l}\text{Note that} \\ x \to x^{(0)} \text{ is the} \\ \text{same as } \|x - x^{(0)}\| \to 0\end{array}\right)$$

If f is differentiable at $x^{(0)}$, v is called the gradient of f at $x^{(0)}$, denoted by $\nabla f(x^{(0)})$.

Example 1: $f(x) = \|x\|^2$, where $\|x\|$ is the norm on V.

At any $x^{(0)} \in V$,

$$\|x\|^2 = \|(x - x^{(0)}) + x^{(0)}\|^2 = \langle (x - x^{(0)}) + x^{(0)}, (x - x^{(0)}) + x^{(0)}\rangle$$
$$= \|x - x^{(0)}\|^2 + \|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle$$

Therefore, $\|x\|^2 - \underbrace{\left(\|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle\right)}_{\downarrow \text{ affine approximation}} = \|x - x^{(0)}\|^2$

So $$\lim_{x \to x^{(0)}} \frac{\left| \|x\|^2 - \|x^{(0)}\|^2 - \langle 2x^{(0)}, x - x^{(0)}\rangle \right|}{\|x - x^{(0)}\|} = \lim_{x \to x_0} \frac{\|x - x^{(0)}\|^2}{\|x - x^{(0)}\|} = 0.$$

Thus, $\nabla f(x^{(0)}) = 2x^{(0)}$.

Example 2: $f(x) = \langle a, x\rangle$ for some $a \in V$.

At any $x^{(0)} \in V$,

$$\langle a, x\rangle = \langle a, x^{(0)}\rangle + \langle a, x - x^{(0)}\rangle$$

Therefore, $\lim_{x \to x^{(0)}} \frac{|\langle a,x\rangle - \langle a, x^{(0)}\rangle - \langle a, x - x^{(0)}\rangle|}{\|x - x^{(0)}\|} = \lim_{x_0 \to x} \frac{0}{\|x - x^{(0)}\|} = 0.$

Thus, $\nabla f(x^{(0)}) = a$.

Example 3: $f(x) = \|x - a\|^2$ for some $a \in V$.

At any $x^{(0)} \in V$,

$$\begin{aligned} f(x) = \|x - a\|^2 &= \|(x^{(0)} - a) + (x - x^{(0)})\|^2 \\ &= \|x^{(0)} - a\|^2 + \|x - x^{(0)}\|^2 + 2\langle x^{(0)} - a, x - x^{(0)}\rangle \\ &= f(x^{(0)}) + \langle 2(x^{(0)} - a), x - x^{(0)}\rangle + \|x - x^{(0)}\|^2 \end{aligned}$$

So, $\lim_{x \to x^{(0)}} \frac{|f(x) - f(x^{(0)}) - \langle 2(x^{(0)} - a), x - x^{(0)}\rangle|}{\|x - x^{(0)}\|} = \lim_{x \to x^{(0)}} \|x - x^{(0)}\| = 0.$

Therefore, $\nabla f(x^{(0)}) = 2(x^{(0)} - a)$

Properties:

① Frechet differentiation is linear $\overset{\inf}{,}$ i.e,

$$\nabla(\alpha f + \beta g)(x) = \alpha \nabla f(x) + \beta \nabla g(x).$$

Proof. By definition,

$$\begin{cases} \lim_{\|y - x\| \to 0} \frac{|f(y) - f(x) - \langle \nabla f(x), y - x\rangle|}{\|y - x\|} = 0 \\ \lim \frac{|g(y) - g(x) - \langle \nabla g(x), y - x\rangle|}{} = 0 \end{cases}$$

$\lim_{\|y-x\|\to 0} \frac{|\quad|}{\|y-x\|} = 0$

Therefore,

$$E = |(\alpha f+\beta g)(y) - (\alpha f+\beta g)(x) - \langle \alpha \nabla f(x) + \beta \nabla g(x), y-x\rangle|$$
$$\le |\alpha| \cdot |f(y) - f(x) - \langle \nabla f(x), y-x\rangle|$$
$$+ |\beta| \cdot |g(y) - g(x) - \langle \nabla g(x), y-x\rangle|$$

and $0 \le \lim_{\|y-x\|\to 0} \frac{E}{\|x-y\|} \le \lim_{\|y-x\|\to 0} \frac{|\alpha|\,|\cdots| + |\beta|\,|\cdots|}{\|y-x\|} = 0$ ∎

② Chain rule: Let $f: V \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$. Then $g\circ f: V \to \mathbb{R}$ and

$$\nabla(g\circ f)(x) = g'(f(x))\,\nabla f(x)$$

if f and g are differentiable at x and $f(x)$ respectively.

Proof. By definition,

$$\lim_{\|y-x\|\to 0} \frac{|f(y)-f(x) - \langle \nabla f(x), y-x\rangle|}{\|y-x\|} = 0 \quad ①$$
$$\lim_{s\to t} \frac{|g(s)-g(t) - \langle g'(t), s-t\rangle|}{|s-t|} = 0 \quad ②$$

choose $s = f(y)$, $t = f(x)$.

① $\Rightarrow$ $f(y) = f(x) + \langle \nabla f(x), y-x\rangle + o(\|y-x\|)$

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

$\le \|\nabla f(x)\|\,\|y-x\| + o(\|y-x\|) \to 0$ as $\|y-x\| \to 0$. ③

Thus,

$$0 \le \lim_{\|y-x\|\to 0} \frac{|\,g(f(y)) - g(f(x)) - g'(f(x))\langle \nabla f(x), y-x\rangle|}{\|y-x\|}$$

$$\le \lim_{\|y-x\|\to 0} \left(\frac{|g(f(y)) - g(f(x)) - g'(f(x))(f(y)-f(x))|}{\|y-x\|} + \frac{|g'(f(x))| \cdot |f(y)-f(x)-\langle \nabla f(x), y-x\rangle|}{\|y-x\|} \right)$$

I_1

I_2

① $\Rightarrow \lim_{\|y-x\|\to 0} I_2 = 0$

Also, $\lim_{\|y-x\|\to 0} I_1 = \lim_{\|y-x\|\to 0} \frac{|g(f(y)) - g(f(x)) - g'(f(x))(f(y)-f(x))|}{|s-t|} \cdot \frac{|s-t|}{\|y-x\|}$

$\leq \lim_{s\to t} \frac{|g(f(y)) - g(f(x)) - g'(f(x))(f(y)-f(x))|}{|s-t|} \cdot \lim_{\|y-x\|\to 0} \frac{|s-t|}{\|y-x\|} = 0$

finite by ③

Example 4: $f(x) = \|x\| \quad \forall\, x \in V.$

This is a composition of $f_1(x) = \|x\|^2$ from $V \to \mathbb{R}$

and $f_2(t) = \sqrt{t}$ from $\mathbb{R} \to \mathbb{R}$.

When $\|x\| \neq 0$, both f_1 and f_2 are differentiable.

Also, $\nabla f_1(x) = 2x$, $f_2'(t) = \frac{1}{2\sqrt{t}}$ if $t \neq 0$.

So, $\nabla f(x) = \nabla (f_2 \circ f_1)(x) = f_2'(f_1(x)) \cdot \nabla f_1(x)$

$$= \frac{1}{2\sqrt{\|x\|^2}} \cdot 2x = \frac{x}{\|x\|}.$$

When $\|x\| = 0$, (i.e., $x = 0$), $f_2(t)$ is NOT differentiable at $f_1(x) = 0$.

It can be shown that $f(x) = \|x\|$ is NOT differentiable at $x = 0$.

③ For functions on $\mathbb{R}^n$: $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}, \quad \text{where} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

i.e., Frechet differentiation is the same as the standard differentiation in multivariate calculus.

Proof. Then choose $y = x + te_i$ for some e_i and $t \in \mathbb{R}$

$$0 = \lim_{\|y-x\|_2 \to 0} \frac{|f(y) - f(x) - \langle \nabla f(x),\, y-x \rangle|}{\|y-x\|_2}$$

$$= \lim_{t \to 0} \frac{|f(x+te_i) - f(x) - \langle \nabla f(x),\, te_i \rangle|}{|t|} = \lim_{t \to 0} \frac{|f(x+te_i) - f(x) - t \cdot (\nabla f(x))_i|}{|t|}$$

Therefore, $(\nabla f(x))_i = \frac{\partial f}{\partial x_i}(x)$ ⊠

Taylor's expansion

From the definition, we see that

$$f(x) \approx f(x^{(0)}) + \langle \nabla f(x^{(0)}),\, x - x^{(0)} \rangle$$

Or, more precisely,

$$f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}),\, x - x^{(0)} \rangle + o(\|x - x^{(0)}\|)$$

This is a generalization of Taylor's expansion.

In particular, if $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) \approx f(x^{(0)}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^{(0)}) \, (x_i - x_i^{(0)}) \, .$$

Differentiation on normed vector spaces

Let V be a normed vector space.

Let $f : V \to \mathbb{R}$ be a function. Let $x^{(0)} \in V$.

To define differentiation, we still use an affine function approximation, and the affine function passes thru $(x^{(0)}, f(x^{(0)}))$.

Thus, we use $f(x^{(0)}) + L(x - x^{(0)})$, where $L: V \to \mathbb{R}$ is a linear function, to approximate $f(x)$.

However, because there is no Riesz representation on normed spaces, we keep the linear function L in the definition of differentiation.

Definition: f is differentiable at $x^{(0)} \in V$, if:

$\exists$ a linear function $L: V \to \mathbb{R}$ such that

$$\lim_{\|x - x^{(0)}\| \to 0} \frac{\left| f(x) - \left(f(x^{(0)}) + L(x - x^{(0)})\right) \right|}{\|x - x^{(0)}\|} = 0 .$$

The linear function L is called the differentiation of f at $x^{(0)}$.

§ 4.4 Case Study: Optimality and Gradient Descent.

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

$$\min_{x \in V} f(x) \qquad (OPT)$$

This section studies the optimality condition for these optimization problems, assuming $f: V \to \mathbb{R}$ is differentiable.

§ 4.4.1. Solvability and Optimality.

- Solvability of (OPT).

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

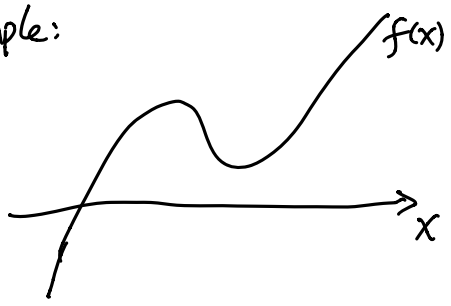
  $$f(x^{(*)}) \leq f(x) \qquad \forall\, x \in V.$$

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

  We also call $x^{(*)}$ a global minimizer of f in V.

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

  Example:

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

  $f(x) = \langle a, x \rangle, \quad a, x \in V.$

  ⋮

  We assume (OPT) has at least one solution.

- Necessary condition for optimality.

  Theorem: Assume f is differentiable at $x^{(*)}$. Then,

  $$x^{(*)} = \arg\min_{x \in V} f(x) \implies \nabla f(x^{(*)}) = 0$$

proof. By expansion,

$$f(x) = f(x^{(*)}) + \langle \nabla f(x^{(*)}),\ x - x^{(*)} \rangle + o\left(\|x - x^{(*)}\|\right)$$

Suppose $\nabla f(x^{(*)}) \neq 0$.

Then choose $\tilde{x} = x^{(*)} - t\,\nabla f(x^{(*)})$ with $t > 0$ gives

$$f(\tilde{x}) = f(x^{(*)}) - t\,\|\nabla f(x^{(*)})\|^2 + o\left(|t|\,\|\nabla f(x^{(*)})\|\right)$$

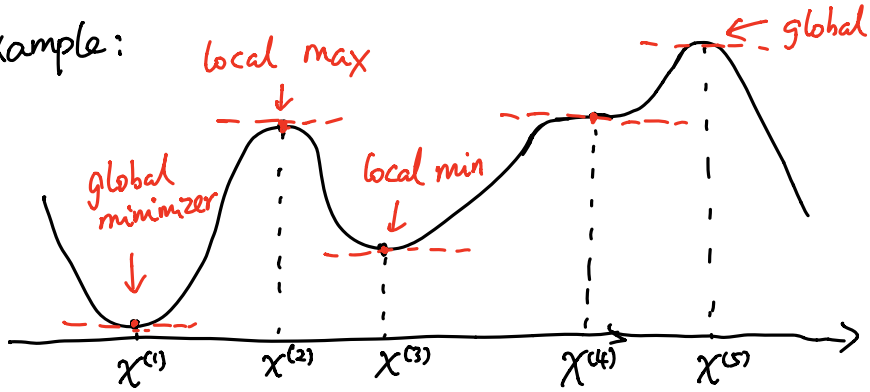
Since $t\|\nabla f(x^{(*)})\| \cdot \|\nabla f(x^{(*)})\|$, by choosing a sufficiently small t,

$$t\,\|\nabla f(x^{(*)})\|^2 > o\left(|t|\,\|\nabla f(x^{(*)})\|\right),$$

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 "$\nabla f(x^{(*)}) = 0 \Rightarrow x^{(*)} = \arg\min_{x \in V} f(x)$" 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.

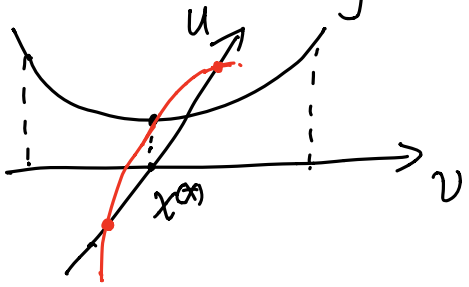
From this example, we see that $x^{(*)}$ with $\nabla f(x^{(*)}) = 0$ can be

- Global minimizer. i.e., $x^{(*)} = \arg\min_{x \in V} f(x)$ (see $x^{(1)}$)
- Local minimizer, i.e.,
  $\exists\, \varepsilon$, s.t. $f(x^{(*)}) \leq f(x)$ $\forall\, x$: $\|x - x^{(*)}\| \leq \varepsilon$. (see $x^{(2)}$)
- global max, i.e., $\forall\, x \in \mathbb{R}^n$, $f(x^{(*)}) \geq f(x)$, (See $x^{(5)}$)
- Local max, i.e, $\exists\, \varepsilon$. s.t. $f(x^{(*)}) \geq f(x)$ $\forall\, x$: $\|x - x^{(*)}\| \leq \varepsilon$
  (See $x^{(3)}$)
- Saddle points (only for V with $\dim(V) \geq 2$), i.e.,

$\exists\, u, v \in \mathbb{R}^n$ s.t. $f(x^{(*)}) \geq f(x^{(*)} + tu)$ for all $|t| \leq \varepsilon$.

and $f(x^{(*)}) \leq f(x^{(*)} + tv)$ for all $|t| \leq \varepsilon$.

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



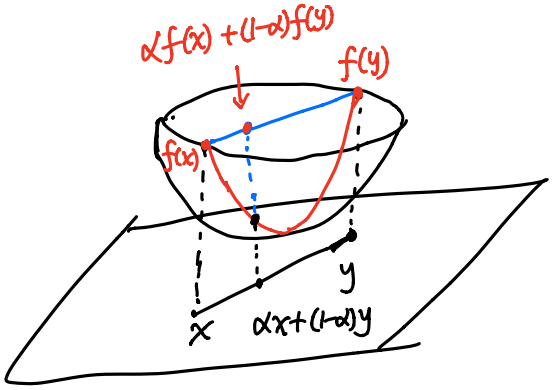
- None of the above (see $x^{(4)}$)

- Sufficient condition for optimality

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

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \forall\, x, y \in V,\ \alpha \in [0,1].$$

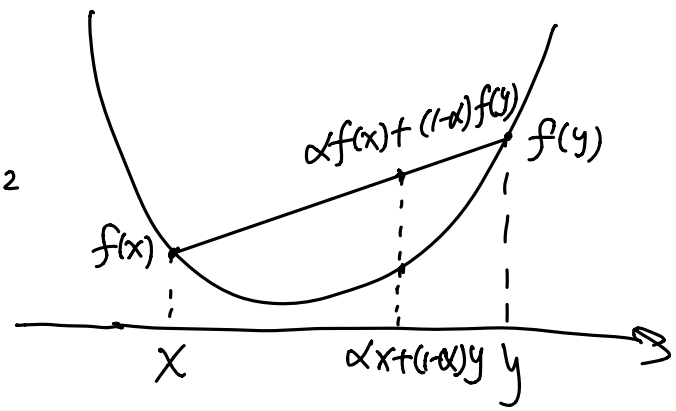
(Sublinearity with $0 \leq \alpha, 1-\alpha \leq 1$)



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

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

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



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

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

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

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

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.

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

Theorem: If f is convex and differentiable, then

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

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 $x \neq y$, and $\alpha \in [0,1]$. Consider $z = \alpha x + (1-\alpha) y$.

Then $f(x) \ge f(z) + f'(z)(x-z)$ (1)

$f(y) \ge f(z) + f'(z)(y-z)$ (2)

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

Next, we prove the general case, i.e, $f: V \to \mathbb{R}$.

"$\Rightarrow$". 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,

$$\begin{aligned} g(\alpha s + (1-\alpha)t) &= f\big((\alpha s + (1-\alpha)t)x + (1 - \alpha s - (1-\alpha)t)y\big) \\ &= f\big(\alpha(sx + (1-s)y) + (1-\alpha)(tx + (1-t)y)\big) \\ &\le \alpha f(sx + (1-s)y) + (1-\alpha) f(tx + (1-t)y) = \alpha g(s) + (1-\alpha) g(t) \end{aligned}$$

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

"$\Rightarrow$". Choose $x \neq y$, $\alpha \in [0,1]$, 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) $\times (1-\alpha)$ $\Rightarrow$ $f(z) \le \alpha f(x) + (1-\alpha) f(y)$. ⊠

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 $x \in \mathbb{R}^n$,

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

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

$$\Rightarrow \quad f(x) \geq f(x^{(*)}), \quad \text{i.e.,} \quad x^{(*)} = \arg\min f(x). \quad \blacksquare$$

§ 4.4.2 Gradient Descent

- Let $f: V \to \mathbb{R}$ be a differentiable function on a Hilbert space V.
- We now consider the numerical solution of

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

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

  We want to find a better estimation $x^{(k+1)}$
- Since f is differentiable, we have

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

  and the approximation is accurate provided $\|x - x^{(k)}\|$ is small.
- Instead finding a global minimizer of $f(x)$, we find a minimizer of $f(x)$ locally in a neighborhood of $x^{(k)}$, i.e.,

$$x^{(k+1)} = \text{solution of} \begin{cases} \min_{x \in V} f(x) \\ \text{s.t.} \quad \|x - x^{(k)}\| \leq \alpha_k \|\nabla f(x^{(k)})\| \end{cases}, \quad \alpha_k > 0 \text{ is a parameter.}$$

  a small $\|\nabla f(x^{(k)})\|$ implies $x^{(k)}$ is close to a solution.

Obviously, this algorithm converges to a local min of f.

However, the algorithm is not practical, because it is difficult to find a solution of the sub-problem.

- By using the linear approximation

$$\begin{cases} \min\limits_{f\in V} f(x) \\ \text{s.t. } \|x-x^{(k)}\| \le \alpha_k \|\nabla f(x^{(k)})\| \end{cases} \overset{\text{approx}}{\Longleftrightarrow} \begin{cases} \min\limits_{x\in V} f(x^{(k)}) + \langle \nabla f(x^{(k)}),\, x-x^{(k)} \rangle \\ \text{s.t.} \quad \|x-x^{(k)}\| \le \alpha_k \|\nabla f(x^{(k)})\| \end{cases}$$

$$\Longleftrightarrow \begin{cases} \min\limits_{x\in V} \langle \nabla f(x^{(k)}),\, x-x^{(k)} \rangle \\ \text{s.t} \quad \|x-x^{(k)}\| \le \alpha_k \|\nabla f(x^{(k)})\| \end{cases} \quad (*)$$

where $\alpha_k > 0$ is a small number.

- The problem $(*)$ can be solved exactly as follows:

— By Cauchy-Schwartz,

For any $x\in V$ satisfying $\|x-x^{(k)}\| \le$, we have

$$\langle \nabla f(x^{(k)}),\, x-x^{(k)} \rangle \ge -\|\nabla f(x^{(k)})\|\,\|x-x^{(k)}\|$$

$$\ge -\alpha_k \|\nabla f(x^{(k)})\|^2 \quad \text{— constant}$$

Therefore, the minimum value of $(*) \ge -\alpha_k \|\nabla f(x^{(k)})\|^2$

— Choose $x\in V$ s.t. $x-x^{(k)} = -\alpha_k \nabla f(x^{(k)})$. Then

- $\|x-x^{(k)}\| = \alpha_k \|\nabla f(x^{(k)})\| \le \alpha_k \|\nabla f(x^{(k)})\|$ — constraint in $(*)$ is satisfied.
- $\langle \nabla f(x^{(k)}),\, x-x^{(k)} \rangle = \langle \nabla f(x^{(k)}),\, -\alpha_k \nabla f(x^{(k)}) \rangle$

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

Therefore, the minimum value of $(*) \le -\alpha_k \|\nabla f(x^{(k)})\|^2$

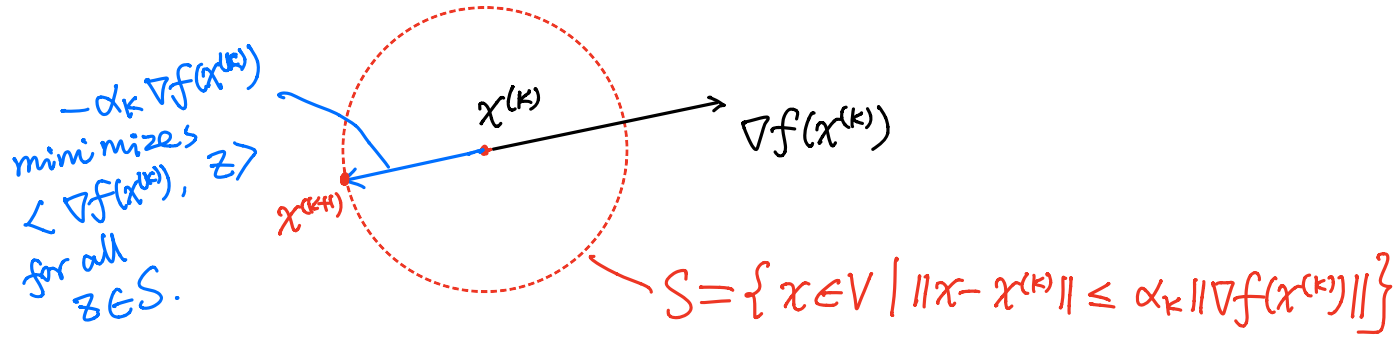
Thus, the minimum value of $(*) = -\alpha_k \|\nabla f(x^{(k)})\|^2$, and the minimum is attained at $x\in V$ s.t. $x-x^{(k)} = -\alpha_k \nabla f(x^{(k)})$,

i.e., $x = x^{(k)} - \alpha_k \nabla f(x^{(k)})$ is the solution of $(*)$.

- So, we obtain

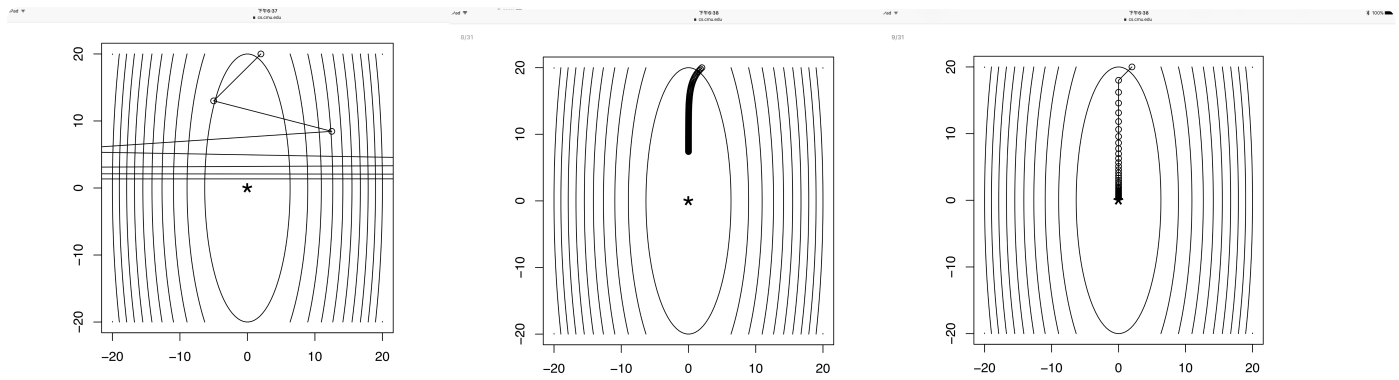
$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}), \quad k = 0, 1, 2, \ldots$$

This algorithm is known as Gradient Descent Algorithm.



- The parameter α_k is called a step size.

  There are many different strategies for choosing α_k.
- Gradient descent converges to a global min of $f(x)$ only if f is convex.
- If f is not convex, the gradient descent finds only a point $\nabla f(x) = 0$.



too large α_k | too small α_k | proper α_k

§ 4.4.3. Examples

① Least Squares:

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

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. To find its gradient, we use the following theorem (a generalization of the chain rule)

> Theorem: Let $g:\mathbb{R}^m\to\mathbb{R}$ be a differentiable function. Let $A\in\mathbb{R}^{m\times n}$ be a matrix. Define $h:\mathbb{R}^n\to\mathbb{R}$ by $h(x)=g(Ax)$.
> Then $\nabla h(x) = A^T\cdot\nabla g(Ax)$

proof. Consider the affine approximation of h:

$$\begin{aligned} E(x,y) &= h(y)-h(x)+\langle A^T\nabla g(Ax),\ y-x\rangle \\ &= h(y)-\big(h(x)+\langle\nabla g(Ax),\ A(y-x)\rangle\big) \\ &= g(Ay)-\big(g(Ax)+\langle\nabla g(Ax),\ Ay-Ax\rangle\big) \end{aligned}$$

Recall in $\mathbb{R}^m, \mathbb{R}^n$
$\langle A^Tu, v\rangle = v^TA^Tu$
$= (Av)^Tu = \langle u, Av\rangle$

We then have

$$\lim_{\|y-x\|_2\to 0} \frac{E(x,y)}{\|y-x\|_2} = \lim_{\|Ay-Ax\|_2\to 0} \frac{E(x,y)}{\|y-x\|_2} \quad \left(\text{Because } \|Ay-Ax\|_2 = \|A(y-x)\|_2 \le \|A\|_2\|y-x\|_2 \to 0\right)$$

$$\le \lim_{\|Ay-Ax\|_2\to 0} \frac{E(x,y)\cdot\|A\|_2}{\|Ay-Ax\|_2} \quad \left(\text{Because } \|Ay-Ax\|_2 \le \|A\|_2\|y-x\|_2\right)$$

$$= \frac{1}{\|A\|_2}\cdot \lim_{\|Ay-Ax\|_2\to 0} \frac{E(x,y)}{\|Ay-Ax\|_2}$$

$$= 0 \quad \text{(Because } g \text{ is differentiable)}$$

Therefore, $\nabla h(x) = A^T \cdot \nabla g(Ax)$ ⊠

By applying the theorem: set $g(y) = \frac{1}{2}\|y-b\|_2^2$,

Then $f(x) = \|Ax-b\|_2^2 = g(Ax)$ and $\nabla g(y) = y-b$.

and $\nabla f(x) = A^T \nabla g(Ax) = A^T(Ax-b)$

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

- Therefore,

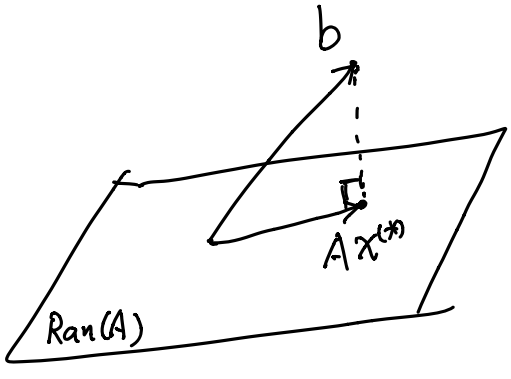
$$x^{(*)} = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 \iff \boxed{A^TAx^{(*)} = A^Tb}$$

↑ 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,

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

which is the same as

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

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

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

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

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

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

Intialize $x^{(0)}$

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

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

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

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

end

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

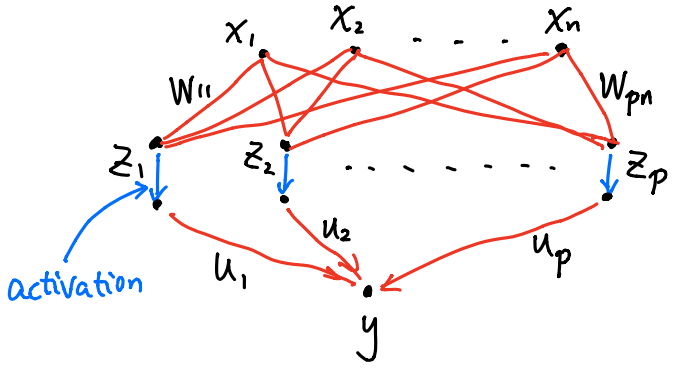
$$f(x^{(i)}) \approx y_i \quad , \quad i = 1, 2, \cdots, m.$$

In linear regression, we choose f be an affine function.

In kernel regression, we choose f be a linear function on the feature space.

In deep learning, we choose f be a function generated from an deep neural network.

For simplicity, we consider a fully-connected neural network with one hidden layer



Input layer $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$

Hidden Layer $z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} \in \mathbb{R}^p$

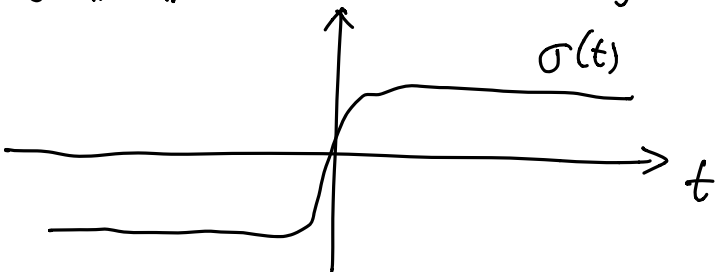
output layer $y \in \mathbb{R}$

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 of the neural network 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



Define $f_{W,u}(x) = \sum_{j=1}^{p} u_j \sigma(\langle w^{(j)}, x \rangle)$

Neural network training
$\iff$ finding W, u s.t. $f_{W,u}(x^{(i)}) \approx y_i \quad i=1, \cdots, m$

For this purpose, we minimize

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

Let $F_i(W,u) = \left(f_{W,u}(x^{(i)}) - y_i \right)^2 = \left(\sum_{j=1}^{p} u_j \sigma(\langle w^{(j)}, x^{(i)} \rangle) - y_i \right)^2$

and $F(W,u) = \sum_{i=1}^{m} F_i(W,u)$.

We need to solve

$$\min_{W,u} F(W,u)$$

- In general, $F(W,u)$ is NOT convex.
- Assume σ is differentiable. Then $F(W,u)$ is differentiable.
  So, we may use gradient descent to solve
  $$\min_{W,u} F(W,u)$$
- Due to the non-convexity,
  $$\min_{W,u} F(W,u) \not\iff \nabla F(W,u) = 0$$
  Therefore, gradient descent doesn't guarantee the global min.
- To perform the gradient descent, let's calculate $\nabla F(W,u)$:
  - Because
    $$F(W,u) = \sum_{i=1}^{m} F_i(W,u),$$
    it suffices to find $\nabla F_i(W,u)$ for all i.
  - Because $W \in \mathbb{R}^{n \times p}$, $u \in \mathbb{R}^p$,
    $$\nabla F_i(W,u) = \left[\frac{\partial}{\partial u_k} F_i(W,u), \ \frac{\partial}{\partial w^{(k)}} F_i(W,u) \right]_{k=1}^{p}$$
    $\frac{\partial}{\partial u_k} F_i(W,u)$:
    $$\frac{\partial}{\partial u_k} F_i(W,u) = \frac{\partial}{\partial u_k} \left(\sum_{j=1}^{p} u_j \sigma(\langle w^{(j)}, x^{(i)} \rangle) - y_i \right)^2$$

$$= 2\left(\sum_{j=1}^{p} u_j\,\sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right)\cdot \frac{\partial}{\partial u_k}\left(\sum_{j=1}^{p} u_j\,\sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right)$$

$$= 2\left(\sum_{j=1}^{p} u_j\,\sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right)\cdot \frac{\partial}{\partial u_k}\left(u_k\,\sigma(\langle w^{(k)}, x^{(i)}\rangle)\right)$$

$$= 2\left(\sum_{j=1}^{p} u_j\,\sigma(\langle w^{(j)}, x^{(i)}\rangle) - y_i\right)\cdot \sigma(\langle w^{(k)}, x^{(i)}\rangle)$$

$$= 2\left(f_{w,u}(x^{(i)}) - y_i\right)\cdot \sigma(\langle w^{(k)}, x^{(i)}\rangle)$$

$\frac{\partial}{\partial w^{(k)}} F_i(w,u)$:

Let $g(w^{(k)}) = \left(u_k\,\sigma(\langle w^{(k)}, x^{(i)}\rangle) + \underbrace{\sum_{j\neq k} u_j\,\sigma(\langle w^{(j)}, x^{(i)}\rangle)}_{=A} - y_i\right)^2$

Let $g_1(t) = (t+A)^2 \qquad g_1 : \mathbb{R}\to\mathbb{R}$

$$g_1'(t) = 2(t+A)$$

Let $g_2(z) = u_k\,\sigma(z), \qquad g_2 : \mathbb{R}\to\mathbb{R}$

$$\Rightarrow \quad g_2'(z) = u_k\,\sigma'(z)$$

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

Note that $g(w^{(k)}) = g_1\left(g_2\left((x^{(i)})^T w\right)\right)$

The chain rule gives

$$\nabla g(w^{(k)}) = (x^{(i)})\cdot (g_1\circ g_2)'((x^{(i)})^T w)$$

$$= g_1'\left(g_2((x^{(i)})^T w)\right)\cdot g_2'((x^{(i)})^T w)\cdot x^{(i)}$$

$$= 2\left(f_{w,u}(x^{(i)}) - y_i\right)\cdot u_k\,\sigma'(\langle w^{(k)}, x^{(i)}\rangle)\cdot \langle x^{(i)}, u\rangle$$

Therefore,

$$\frac{\partial}{\partial w^{(k)}} F_i(w,u) = 2\left(f_{w,u}(x^{(i)}) - y_i\right)\cdot u_k\,\sigma'(\langle w^{(k)}, x^{(i)}\rangle)\cdot x^{(i)}$$

- So, Gradient descent for nueral network training

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

$$\begin{cases} u_k^{(l+1)} = u_k^{(l)} - \alpha_l\left(\sum_{i=1}^{m} \frac{\partial}{\partial u_k} F_i(w,u)\right) \\ w^{(k,l+1)} = w^{(k,l)} - \alpha_l\left(\sum_{i=1}^{m} \frac{\partial}{\partial w^{(k)}} F_i(w,u)\right) \end{cases} \qquad k = 1, 2, \cdots, p$$

end

- In practics, m is very large, Therefore, the gradient descent may be too expensive to find ∇F
  Then, a popular algorithm is Stochastic Gradient Descent, where the following strategy is used:
  - Assume the training data $\{x_i, y_i\}_{i=1}^m$ are independent identically distributed (i.i.d.) at random. Then $F_i(W,u)$, $i=1,\cdots,m$ are i.i.d. random functions. Intuitively, every $\nabla F_i(W,u)$, $i=1,\cdots,m$ should be the same statistically. Therefore,

$$\nabla F(W,u) = \sum_{i=1}^{m} \nabla F_i(W,u) \underset{\text{statistically}}{\sim} m \nabla F_i(W,u) \text{ for any } i.$$

  At each step l:
  — Choose a random $i_l \in \{1,2,\cdots,m\}$.
  — Use $m\nabla F_{i_l}(\cdot,\cdot)$ to approximate $\nabla F(\cdot,\cdot) = \sum_{i=1}^{m} \nabla F_i(\cdot,\cdot)$

for $l = 0, 1, 2, \cdots$
  Choose a traing example i_l from $\{1,2,\cdots,m\}$ randomly.

$$u_k^{(l+1)} = u_k^{(l)} - \alpha_l \frac{\partial}{\partial u_k} F_{i_l}(W,u) \qquad k=1,2,\cdots,p$$

$$W^{(k,l+1)} = W^{(k,l)} - \alpha_l \frac{\partial}{\partial W^{(k)}} F_{i_l}(W,u)$$

end