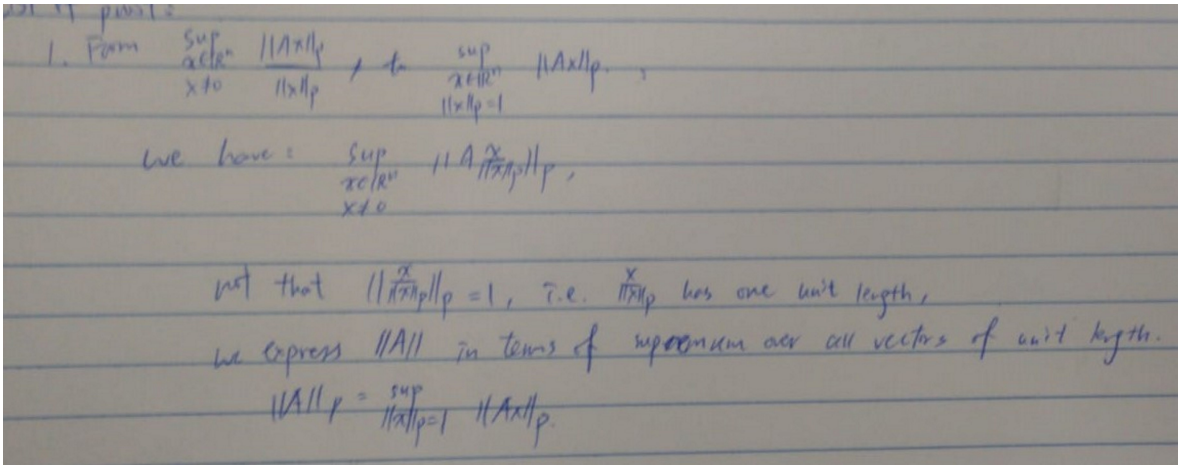
Final Review - Proofs

Sunday, 6 December 2020 2:18 AM

$$\sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \Rightarrow \sup_{x \in \mathbb{R}^n, \|x\|_p = 1} \|Ax\|_p.$$



2. Prove operator matrix norm is indeed a norm: $\|A\|_{p \to p}$

(P2) First consider $\mathbf{A} = \mathbf{0}$, the all-zero matrix. Clearly,

$$\|\mathbf{0}\|_{a,b} = \sup_{\|\mathbf{x}\|_b \leq 1} \|\mathbf{0x}\|_a = 0 \tag{11}$$

Next suppose that $\|\mathbf{A}\|_{a,b} = 0$ for some $\mathbf{A}$. Then,

$$\sup_{\|\mathbf{x}\|_b \leq 1} \|\mathbf{Ax}\|_a = 0 \tag{12}$$

$$\Rightarrow \|\mathbf{Av}\|_a = 0 \qquad \text{for all } \|\mathbf{v}\|_b \leq 1, \text{ by definition} \tag{13}$$

$$\Rightarrow \mathbf{Av} = 0 \qquad \text{from (P2), for all } \|\mathbf{v}\|_b \leq 1 \tag{14}$$

$$\Rightarrow \mathbf{A} = \mathbf{0} \tag{15}$$

The last step follows since, if $\mathbf{A}$ has even a single non-zero entry, it will always be possible to find an $\mathbf{v} : \|\mathbf{v}\|_b \leq 1$, such that $\mathbf{Av} \neq 0$.

(P3) Trivial, since

$$\sup_{\|\mathbf{x}\|_b \leq 1} \|t\mathbf{Ax}\|_a = \sup_{\|\mathbf{x}\|_b \leq 1} |t| \, \|\mathbf{Ax}\|_a = |t| \sup_{\|\mathbf{x}\|_b \leq 1} \|\mathbf{Ax}\|_a \tag{16}$$

from (P3) for vector norm.

(P4) Consider the rhs

$$\|\mathbf{A}\|_{a,b} + \|\mathbf{B}\|_{a,b} = \sup_{\|\mathbf{x}\|_b \leq 1} \|\mathbf{Ax}\|_a + \sup_{\|\mathbf{y}\|_b \leq 1} \|\mathbf{By}\|_a \qquad \text{by definition} \tag{17}$$

$$\geq \|\mathbf{Av}\|_a + \|\mathbf{Bu}\|_a \qquad \forall \ \|\mathbf{v}\|_b \leq 1, \|\mathbf{u}\|_b \leq 1 \tag{18}$$

This statement holds for all $\mathbf{u}$, $\mathbf{v}$ such that $\|\mathbf{v}\|_b \leq 1, \|\mathbf{u}\|_b \leq 1$. Therefore it also holds when $\mathbf{u} = \mathbf{v}$ with $\|\mathbf{u}\|_b \leq 1$, i.e.,

$$\|\mathbf{A}\|_{a,b} + \|\mathbf{B}\|_{a,b} \geq \|\mathbf{Au}\|_a + \|\mathbf{Bu}\|_a \qquad \forall \ \|\mathbf{u}\|_b \leq 1 \tag{19}$$

$$\geq \|\mathbf{Au} + \mathbf{Bu}\|_a \qquad \forall \ \|\mathbf{u}\|_b \leq 1 \quad \text{from (P4)} \tag{20}$$

which implies that

$$\|\mathbf{A}\|_{a,b} + \|\mathbf{B}\|_{a,b} \geq \sup_{\|\mathbf{u}\|_b \leq 1} \|\mathbf{Au} + \mathbf{Bu}\|_a \tag{21}$$

$$= \|\mathbf{A} + \mathbf{B}\|_{a,b} \tag{22}$$

$$= \|\mathbf{A} + \mathbf{B}\|_{a,b} \tag{22}$$

3. Proof the following computation of l-inf norm, l1 and l2-norm

Example: $a = \left(1, \frac{1}{2}, \frac{1}{3}, \dots\right)$, $\|a\|_\infty = 1$, $\|a\|_2 = \frac{\pi}{6}$, $\|a\|_1 = +\infty$, 4

$\therefore a \in l_\infty$, $a \in l_2$, $a \notin l_1$ 1 of 3

Proof:

Then $\|a\|_\infty = \sup_i |a_i| = 1$ 呢個係從sin(x) taylor series推的

$$\|a\|_2 = \left(\sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^2\right)^{1/2} = \left(\frac{\pi^2}{6}\right)^{1/2} = \frac{\pi}{\sqrt{6}}$$

$$\|a\|_1 = \left(\sum_{i=1}^{\infty} \frac{1}{i}\right) = +\infty$$

4. Proof of k-mean algorithms:

Step 1: Fix the representatives $z_1, \dots, z_K$, find the best partitions $G_1, \dots, G_K$, i.e., solve

$$\min_{G_1, \dots, G_K} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \|x_i - z_j\|_2^2\right) \quad \cdots (1)$$

$$(1) \Leftrightarrow \min_{C_1, \dots, C_N} \sum_{j=1}^{K} \sum_{i \in G_j} \|x_i - z_j\|_2^2$$ rewrite

finding the partition $G_1, \dots, G_K$ is equivalent to finding $C_1, C_2, \dots, C_N$. So (1) becomes

$$\min_{C_1, C_2, \dots, C_N} \left(\underbrace{\|x_1 - z_{C_1}\|_2^2}_{\text{depends on } C_1 \text{ only}} + \underbrace{\|x_2 - z_{C_2}\|_2^2}_{\text{depends on } C_2 \text{ only}} + \cdots + \underbrace{\|x_N - z_{C_N}\|_2^2}_{\text{depends on } C_N \text{ only}}\right)$$

$$\Updownarrow$$

$$\min_{C_i} \|x_i - z_{C_i}\|_2^2 \qquad i = 1, 2, \dots, N.$$

Since $C_i \in \{1, 2, \dots, k\}$, to get C_i, we only need to compare

$$\|x_i - z_1\|_2^2, \|x_i - z_2\|_2^2, \dots, \|x_i - z_K\|_2^2$$

and choose the minimum from it. i.e.,

$$C_i = \arg\min_{j \in \{1, \dots, K\}} \|x_i - z_j\|_2^2, \quad i = 1, 2, \dots, K.$$

In other words,

x_i is assigned to the group whose representative vector is the closest to x_i.

Step 2: Fix the groups $G_1, \cdots, G_K$, find the best representatives $z_1, \cdots, z_K$, i.e., solve

$$\min_{z_1,\cdots,z_K} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \|x_i - z_j\|_2^2 \right) \quad ---- \quad (2)$$

$$(2) \Leftrightarrow \min_{z_1,\cdots,z_K} \sum_{j=1}^{K} \sum_{i \in G_j} \|x_i - z_j\|_2^2$$

Swap of summation order

$$\Leftrightarrow \min_{z_1,\cdots,z_K} \left(\sum_{i \in G_1} \|x_i - z_1\|_2^2 + \sum_{i \in G_2} \|x_i - z_2\|_2^2 + \cdots + \sum_{i \in G_K} \|x_i - z_K\|_2^2 \right)$$

$$\Leftrightarrow \min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_2^2, \quad j = 1, \cdots, K$$

independent

Taking derivative w.r.t. z_{jl} and setting it to 0, we obtain that the solution z_{jl} satisfies

$$2 \sum_{i \in G_j} (z_{jl} - x_{il}) = 0$$

$$\Rightarrow z_{jl} = \left(\sum_{i \in G_j} x_{il} \right) / |G_j| \quad (|G_j| \text{ is the number of elements in } G_j)$$

$l = 1, 2, \cdots, n.$

In vector form,

$$\begin{pmatrix} z_{j1} \\ z_{j2} \\ \vdots \\ z_{jn} \end{pmatrix} = \frac{1}{|G_j|} \cdot \sum_{i \in G_j} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix} \Leftrightarrow z_j = \frac{1}{|G_j|} \left(\sum_{i \in G_j} x_i \right)$$

In other words,

z_j is the mean of all vectors in G_j.

5.

Lemma: If $x^{(k)} \to x \in V$, then $\{x^{(k)}\}$ is a cauchy sequence.

Proof. $x^{(k)} \to x$ implies: $\forall \frac{\varepsilon}{2} > 0$, $\exists K$ s.t. $\forall k \geq K$ $\|x^{(k)} - x\| \leq \frac{\varepsilon}{2}$.

Therefore, $\|x^{(k)} - x^{(l)}\| \leq \|x^{(k)} - x\| + \|x^{(l)} - x\| \leq \varepsilon$. $\forall k, l \geq K$. ⊠

6. If $\{x^{(k)}\}$ is a cauchy sequence, $x^{(k)} \to x \in V$ may not hold.

Consider this example:

Let $a^{(k)} = \begin{pmatrix} 1 \\ 1/2 \\ \vdots \\ 1/k \\ 0 \\ \vdots \end{pmatrix}$ $a = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ \vdots \\ 1/k \\ \vdots \end{pmatrix}$

$\|a^{(k)} - a^{(l)}\|_\infty = \left\| \begin{bmatrix} 0 \\ \vdots \\ 1/k \\ \vdots \\ 1/l \\ 0 \\ \vdots \end{bmatrix} \right\|_\infty = \frac{1}{k} < \varepsilon$ if $k, l \geq \frac{1}{\varepsilon} + 1$ (assume $l \geq k$)

$\Rightarrow \{a^{(k)}\}$ is Cauchy sequence.

$\|a^{(k)}\|_1 = \sum_{i=1}^{k} \left(\frac{1}{i}\right) < +\infty$ as $k < +\infty$.

$\|a\|_1 = \sum_{i=1}^{k} \left(\frac{1}{i}\right) + \sum_{i=1}^{\infty} \left(\frac{1}{k}\right) = +\infty$.

$\lim_{k\to\infty} \|a^{(k)} - a\|_\infty = \lim_{k\to\infty} \left\| \begin{pmatrix} 0 \\ \vdots \\ 1/(k+1) \\ 1/(k+1) \\ \vdots \end{pmatrix} \right\|_\infty = \lim_{k\to\infty} \frac{1}{k+1} = 0$

$a^{(k)} \to a$, $a^k \in V$, but $a \notin V$

$\therefore$ It is proved that if $\{x^{(k)}\}$ is Cauchy sequence,

$x^{(k)} \to x \in V$.

7.

Note that the constants C_1, C_2 depend on V.

Example: Consider $\mathbb{R}^n$ and $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$

- $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent because
$$\|a\|_2 \leq \|a\|_1 \leq \sqrt{n}\,\|a\|_2$$
- $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equivalent
$$\|a\|_\infty \leq \|a\|_2 \leq \sqrt{n}\,\|a\|_\infty$$
- $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are equivalent
$$\|a\|_\infty \leq \|a\|_1 \leq n\,\|a\|_\infty$$

TBC....

8. Proof of Cauchy Schwartz Inequality

Proof • if $y=0$. Then obviously

$$|\langle x,y\rangle|^2 = |\langle x,0\rangle|^2 = 0 \le 0 = \langle x,x\rangle\langle y,y\rangle$$

• It remains to prove the inequality with $y \neq 0$.

Let $\lambda \in \mathbb{R}$ be an arbitrary number

$$0 \le \langle x+\lambda y, x+\lambda y\rangle = \langle x,x\rangle + \lambda\langle y,x\rangle + \lambda\langle x,y\rangle + \lambda^2\langle y,y\rangle$$

$$= \langle x,x\rangle + 2\lambda\langle x,y\rangle + \lambda^2\langle y,y\rangle$$

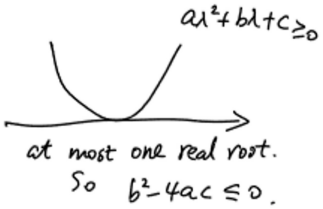
Thus, $\lambda^2\langle y,y\rangle + 2\lambda\langle x,y\rangle + \langle x,x\rangle \ge 0$. $\forall \lambda \in \mathbb{R}$.

$= f(\lambda)$

Since $y \neq 0$, $\langle y,y\rangle > 0$.

So, $f(\lambda)$ a quadratic function of λ that takes non-negative values only.

There is at most one root of the quadratic function.



So, $(2\langle x,y\rangle)^2 - 4\langle y,y\rangle\langle x,x\rangle \le 0$

$\Rightarrow \langle x,y\rangle^2 \le \langle x,x\rangle\langle y,y\rangle$

9. Conditions for <x,y>^2 = <x,x><y,y> : x=alpha*y or y = alpha * y

• Next when the equality holds true, i.e., $\langle x,y\rangle^2 = \langle x,x\rangle\langle y,y\rangle$,

  • if $y=0$, then obviously $y=\alpha x$, where $\alpha=0$.

  • if $y\neq 0$, there is exacty one real root of $f(\lambda)$.

    $\exists$ a unique $\lambda \in \mathbb{R}$, $\lambda^2\langle y,y\rangle + 2\lambda\langle x,y\rangle + \langle x,x\rangle = 0$

$\Rightarrow \langle x+\lambda y, x+\lambda y\rangle = 0 \Rightarrow x+\lambda y = 0 \Rightarrow x = \alpha y$ with $\alpha = -\lambda$.

Finally, we show that if $x=\alpha y$ or $y=\alpha y$, the $\langle x,y\rangle^2 = \langle x,x\rangle\langle y,y\rangle$.

10. ||x|| = sqrt(<x,x>) defines a norm

Proof. ① $\|x\| = (\langle x,x\rangle)^{1/2} \ge 0$ and $\|x\| = (\langle x,x\rangle)^{1/2} = 0 \Leftrightarrow x = 0$.

② $\|\alpha x\| = (\langle \alpha x, \alpha x\rangle)^{1/2} = (\alpha^2\langle x,x\rangle)^{1/2} = |\alpha|\,\|x\|$

③ $\|x+y\|_2^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle$

$= \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle$

$\le \|x\|^2 + \|y\|^2 + 2\|x\|\,\|y\|$ (Note that Cacuchy-Schwarts becomes $|\langle x,y\rangle| \le \|x\|\,\|y\|$) ⊗

$= (\|x\| + \|y\|)^2$

11. L2 space is a hilbert space
https://math.stackexchange.com/questions/1259364/show-that-l2-is-a-hilbert-space
12. Rewrite k-mean using kernel trick to map it to hilbert space

第一步：消除 c_j.

簡化成：

① $c_i = \underset{j=1\cdots k}{\operatorname{argmin}} \left\| \phi(x_i) - \frac{1}{|G_j|}\left(\sum_{i\in G_j}\phi(x_i)\right)\right\|$

② $G_j = \{i \mid c_i = j\}$, $j=1\cdots k$.

$$c_i = \underset{j=1\cdots k}{\operatorname{argmin}} \left\|\phi(x_i) - \frac{1}{|G_j|}\left(\sum_{l\in G_j}\phi(x_l)\right)\right\|^2$$

第二步，

加個 square 上去，無影响，same minimizer.

$$\Downarrow$$

$$c_i = \underset{j=1\cdots k}{\operatorname{argmin}} \left\langle \phi(x_i) - \frac{1}{|G_j|}\left(\sum_{l\in G_j}\phi(x_l)\right),\ \phi(x_i) - \frac{1}{|G_j|}\sum_{l\in G_j}\phi(x_l)\right\rangle$$

$$\Updownarrow$$

$$c_i = \underset{j=1\cdots k}{\operatorname{argmin}} \langle\phi(x_i),\phi(x_i)\rangle - 2\left\langle\phi(x_i), \frac{1}{|G_j|}\left(\sum_{l\in G_j}\phi(x_l)\right)\right\rangle + \left\langle\frac{1}{|G_j|}\sum_{l\in G_j}\phi(x_l), \frac{1}{|G_j|}\sum_{l\in G_j}\phi(x_l)\right\rangle$$

$$\Downarrow$$

$$c_i = \underset{j=1\cdots k}{\operatorname{argmin}} \langle\phi(x_i),\phi(x_i)\rangle - \frac{2}{|G_j|}\sum_{l\in G_j}\langle\phi(x_i),\phi(x_l)\rangle + \frac{1}{|G_j|^2}\sum_{l_1\in G_j}\sum_{l_2\in G_j}\langle\phi(x_{l_1}),\phi(x_{l_2})\rangle$$

13. Prove requirements for kernel function

Prove K的一些必要條件:

① $K(x,y) = \langle \phi(x), \phi(y) \rangle \overset{\text{inner product property}}{=} \langle \phi(y), \phi(x) \rangle = K(y,x)$

So, $k(x,y) = k(y,x)$ $\forall x, y \in \mathbb{R}^n$ ⇒ 所以K係 symmetric 的

② Let $y_1, y_2, \cdots, y_m$ 係 $\mathbb{R}^n$ space 中 m 支 vector, for any $c = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} \in \mathbb{R}^m$

$$0 \leq \langle \underbrace{\sum_{i=1}^{m} c_i \phi(y_i)}_{\text{in H space 的, 因為有 } c_i \text{ 呢個實數}}, \sum_{i=1}^{m} c_i \phi(y_i) \rangle = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} c_{i_1} c_{i_2} \langle \phi(y_{i_1}), \phi(y_{i_2}) \rangle$$

(By inner product property)

$$= \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} c_{i_1} c_{i_2} k(y_{i_1}, y_{i_2})$$

$$= \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} c_{i_1} c_{i_2} K_{i_1 i_2}$$

$$= C^T K C$$

$$\Rightarrow C^T K C \geq 0$$

$\therefore K$ is SPSD

⇒ where $K = \begin{bmatrix} k(y_1,y_1) & \cdots & k(y_1,y_m) \\ k(y_2,y_1) & & \vdots \\ \vdots & & \\ k(y_m,y_1) & \cdots & k(y_m,y_m) \end{bmatrix} \in \mathbb{R}^{m \times m}$

⇒ 個係呢個 matrix 裝住

所以K是SPSD係其中一個必要條件

14. Any linear function in R^n can be written as <a,x>

To see this, let $e_1, e_2, \cdots, e_n$, where $e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ ← 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

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

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

15. The inner product for a linear product is unique

- 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 $a = b$.

16. a fixed t0 function cannot have h(t) in <f,h>

$F : C[a,b] \to \mathbb{R}$ defined by $F(f) = f(t_0)$ for a fixed $t_0 \in \mathbb{R}$.

F is linear on $C[a,b]$

If we define an inner product on $C[a,b]$

by $\langle f,g \rangle = \int_a^b f(t)g(t)dt$

and the induced norm is $\|f\|_2 = \left(\int_a^b f(t)^2 dt\right)^{\frac{1}{2}}$

If we want to represent F into inner product form

$f(t_0) = F(f) = \int_a^b f(t)h(t)dt$

we require $h(t) = 0$ when $t \neq t_0$.

But this will make make $\int_a^b f(t)h(t)dt = 0$

We cannot find h.

呢到意思係，我地要令當t不是t_0的時候，
contribute到integral嘅其他term都係0
所以h(t)要係0 when t \neq t_0

17. $S_{a,0} = \{x \in V \mid \langle a,x \rangle = 0\}$ is a hyperplane.

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.

linear

18. $\forall b$, $S_{a,b} = \{x \in V \mid \langle a,x \rangle = b\}$ is also a hyperplane.

Now let's consider $a \in H$

$S_{a,b} = \{x \mid \langle a,x \rangle = b\}$ for $b \in \mathbb{R}$ 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: $b - b = 0$

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

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

any element in $S_{a,b}$ can be written in $S_{a,0} + x_0$, so it is subset

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

In other words,

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

19. Proof of this theorem in projection onto hyperplanes

19. Proof of this theorem in projection onto hyperplanes

Theorem: z is a solution of $\min_{x\in S} \|x-y\|$ if and only if $z\in S$ and $\langle z-y, z-x\rangle = 0 \quad \forall x\in S$.

Proof. ① We first prove that: If $z\in S$ is a solution of $\min_{x\in S} \|x-y\|$, then $\langle z-y, z-x\rangle = 0 \quad \forall x\in S$. arbitrary x

係z與x的那條直線

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, $(1+t)z - tx \in S$.

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

- If we choose $t>0$,
  $\langle z-y, z-x\rangle \ge -\frac{t}{2}\|z-x\|^2$
  Letting $t\to 0_+$ gives $\langle z-y, z-x\rangle \ge 0$.
- If we choose $t<0$,
  $\langle z-y, z-x\rangle \le -\frac{t}{2}\|z-x\|^2$
  Letting $t\to 0_-$ gives $\langle z-y, z-x\rangle \le 0$

$\langle z-y, z-x\rangle$ 只能是0，因為t是任選的

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

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

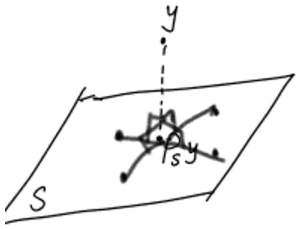
$$\|x-y\|^2 = \|(z-x)-(z-y)\|^2$$

since $\langle z-y, z-x \rangle \geq 0 \quad \forall x \in S$,

$$\begin{aligned} \|x-y\|^2 &= \|(z-x)-(z-y)\|^2 \\ &= \|z-x\|^2 + \|z-y\|^2 - 2\langle z-x, z-y\rangle \\ &= \|z-y\|^2 + \|z-x\|^2 \geq \|z-y\|^2 \quad \forall x \in S. \end{aligned}$$

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

⊗



← solution of $P_S y$ should $\perp$ to all $x \in S$.

20. $P_s y$ should be the following formula:

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 $z \in S$.

② $\forall x \in S$,

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

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

(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, $\Rightarrow \langle z_1 - y, z_1 - z_2\rangle = 0$

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

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

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

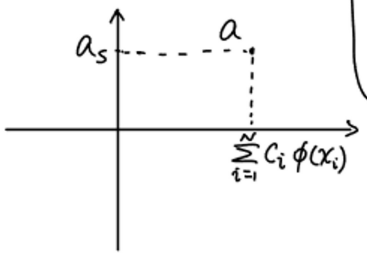
wadiu

21. Proof of representer theorem

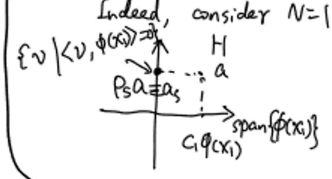
Proof. For any $a \in H$, we claim that a can be decomposed as

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

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 $i = 1, 2, \cdots, N$.



Indeed, consider $N=1$ for simplicity.



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

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, (⊥)

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

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

Also, by direct calculation,

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

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

So, $a = \underbrace{P_S a}_{a_S} + c_1 \phi(x_1)$

For general N, it can be done similarly.

解釋：

a係H space上面

然後將a拆成兩個component

一個係linear combination of phi(x)

一個係與其對應的orthogonal vector

然後我claim that a = 兩個component總和

22. prove how to change objective function to inner product only in kernel regression problem

Therefore,

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

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

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

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

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

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

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

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

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

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

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

Thus, the solution of the original minimization is

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

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

Pf23. SVM-1的推論過程

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

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

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

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

因為x-係S-到，<a,x->自然符合S-定義 = -1-b

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

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

Pf24. SVM-2的邏輯推論

這裡的邏輯是這樣的:

如果 $y_i(\langle a, x_i\rangle + b) - 1 \geq 0$，那麼 $-\xi_i \leq 0$ 同時 fulfill $\xi_i \geq 0$，

$\min \xi_i$ 的話只能是 0.

如果 $-\xi_i \leq y_i(\langle a, x_i\rangle + b) - 1 < 0$，$-\xi_i$ 要細過某個負數，(e.g. -1.5)

同時 $\xi_i \geq 0$. $\Rightarrow \xi_i$ 只能是那負數之絕對值 (1.5)

Pf25: Kernel SVM 推論過程

First step of proof:

We know that: $\forall\, a \in H$, we can decompose it as

$$a = a_S + \sum_{i=1}^{N} c_i \phi(x_i), \quad \text{where } \langle a_S, \phi(x_i)\rangle = 0 \quad i = 1, \cdots, N.$$

Then

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

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

$$= \underbrace{\frac{\lambda}{2}\|a_S\|_H^2}_{F_1(a_S) \Rightarrow a_S = 0} + \underbrace{\left(\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle \phi(x_i), \phi(x_j)\rangle + \sum_{i=1}^{N} h\left(y_i \sum_{j=1}^{N} c_j \langle \phi(x_j), \phi(x_i)\rangle - 1\right)\right)}_{F_2(c)}$$

$c^T K c$ $\qquad$ $h(y_i (Kc)_i - 1)$

Then $\min_{a \in H} F(a)$

$\Updownarrow$

$$\min_{\substack{a_S \in H \\ c \in \mathbb{R}^N \\ \langle a_S, \phi(x_i)\rangle = 0 \\ i = 1, \cdots, N}} F_1(a_S) + F_2(c) \iff \begin{cases} a_S = 0 \\ c \text{ is the solution of } \min_{c \in \mathbb{R}^N} F_2(c) \end{cases}$$

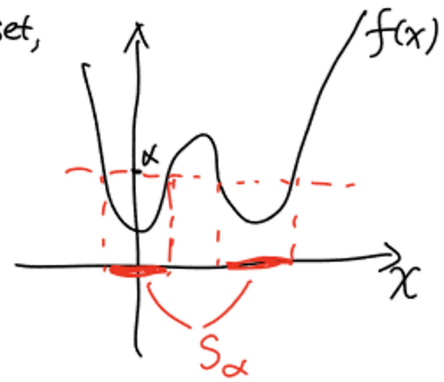
So, a solution of $\min_{a \in H} F(a)$ is

$$a = 0 + \sum_{i=1}^{N} c_i \phi(x_i), \quad \text{where } c \text{ is the solution of } \min_{c \in \mathbb{R}^N} F_2(c).$$

26. Proving if f is continuous and coercive, there exists at least one solution of unconstrained smooth optimization

Proof. Consider the set, known as the level set,

$$S_\alpha = \{ x \mid f(x) \le \alpha \}$$



S_α is closed, because of ①:

Let $x^{(n)} \to x$ and $\{x^{(n)}\} \subset S_\alpha$.
Then $f(x^{(n)}) \le \alpha$
① $\Rightarrow f(x) \le \alpha \Rightarrow x \in S_\alpha$

S_α is bounded, because of ②:

Suppose S_α is unbounded, i.e., $\exists \{x^{(n)}\} \subset S_\alpha$ s.t. $\|x^{(n)}\| \to +\infty$.
Then ② $\Rightarrow f(x^{(n)}) \to +\infty$
but $f(x^{(n)}) \le \alpha$
} contradiction.

Therefore, S_α is closed and bounded. for any $\alpha \in \mathbb{R}$.
We choose α s.t. S_α is non-empty. (For example, choose $\alpha = f(0)$)
By Weierstrass's theorem (any continous function on a bounded and closed non-empty set must have a minimizer and maximizer)
So, $\min_{x \in S_\alpha} f(x)$ exists a solution, called $x^{(*)}$.

Then, $\forall x \in S_\alpha$, $f(x^{(*)}) \le f(x)$.
$\forall x \notin S_\alpha$, $f(x^{(*)}) \le \alpha < f(x)$
($x^* \in S_\alpha$, $x \notin S_\alpha$)
} $\Rightarrow f(x^{(*)}) \le f(x)$ $\forall x \in \mathbb{R}^n$.

⊠

27. proof of first order condition in unconstrained smooth optimization problem

Proof. By definition of the gradient,

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

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

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

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

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

$$\lim_{t \to 0} \frac{t\,\|\nabla f(x^{(*)})\|^2}{t} = \|\nabla f(x^{(*)})\|^2 \neq 0$$

$$\lim_{t \to 0} \frac{o\left(t\,\|\nabla f(x^{(*)})\|\right)}{t} = 0$$

So, by choosing a small enough t,

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

This implies $f(\tilde{x}) < f(x^{(*)})$ contradition.

28. proof of the theorem: Convex + differentiable, grad = 0 <==> argmin

Proof. "$\Longrightarrow$" is proved. (for any f, convex or not)

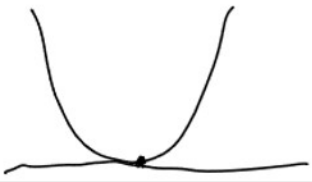
"$\Longleftarrow$"

Since f is convex and differentiable,

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

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

$$\Longrightarrow \quad f(x) \geq f(x^{(*)}) \quad \forall\, x \in \mathbb{R}^n$$



Pf29: Theorem If a function is strictly convex, the solution is unique if it exists

Proof. Suppose there are two distinct solutions $x^{(*)}, y^{(*)}$ of $\min_{x\in\mathbb{R}^n} f(x)$ ($\Rightarrow f(x^{(*)}) = f(y^{(*)})$)

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

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

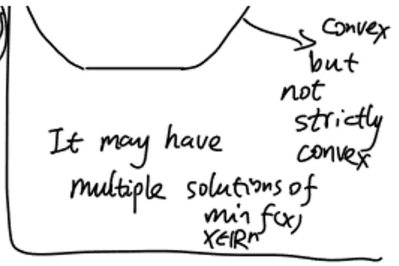
$f(z) = f(\alpha x^{(*)} + (1-\alpha) y^{(*)})$

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

$= f(x^{(*)}) \Rightarrow f(x^{(*)}) > f(z)$

f is strictly convex

$x^{(*)}, y^{(*)} = \arg\min_{x\in\mathbb{R}^n} f(x) \Rightarrow f(x^{(*)}) \leq f(z)$ } contradiction



Pf30 It can be checked F(x) is continuous, coercive, strictly convex

① F is continous, and coercive.

— Continuity is obvious

— Coercivity:

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

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

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

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

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.

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

↑ strictly convex ↑ convex

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

Pf31: proving gradient descent(cont'd)

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,

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

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

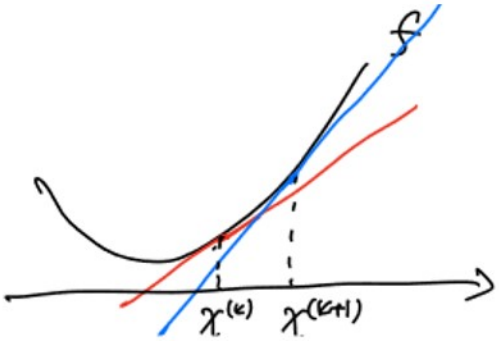
Pf32: Proving range of alpha needs to be(1,m)

For simplicity, we assume f is convex.

By convexity,

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

To give an upper bound of $f(x^{(k+1)})$



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

i.e.,

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

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

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

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

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

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

$$\alpha^2 \|\nabla f(x^{(k)})\|_2^2$$

Assume

$$\leq f(x^{(k)}) - \alpha(1 - M\alpha) \|\nabla f(x^{(k)})\|_2^2$$

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

If we want $f(x^{(k+1)}) < f(x^{(k)})$, we require

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

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

Pf33. objective function for Least Squares is differentiable $\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2$

– f is differentiable, because:

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

$$f(y) = \tfrac{1}{2} \|Ay - b\|_2^2 = \tfrac{1}{2} \|(Ax - b) + A(y - x)\|_2^2$$

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

$$f(y) = \frac{1}{2}\|Ay-b\|_2^2 = \frac{1}{2}\|(Ax-b)+A(y-x)\|_2^2$$

$$= \frac{1}{2}\|Ax-b\|_2^2 + \langle Ax-b, A(y-x)\rangle + \frac{1}{2}\|A(y-x)\|_2^2$$

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

↑ error

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

$$\le \frac{1}{2}\cdot\underbrace{\left(\max_{\|z\|_2\le 1}\|Az\|_2^2\right)}_{\| \\ M}\cdot\|y-x\|_2^2 = \frac{1}{2}M\|y-x\|_2^2$$

So, if M is finite, then

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

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

Indeed, $M = \max_{\|z\|_2\le 1}\|Az\|_2^2 = \left(\max_{\|z\|\le 1}\|Az\|_2\right)^2 = \|A\|_{2\to 2}^2 = \|A\|_2^2$ is finite

Because, $g(z) = \|Az\|_2 = \left(\sum_i\left(\sum_j a_{ij}z_j\right)^2\right)^{1/2}$ is continuous

- $\{z \mid \|z\|_2 \le 1\}$ is closed, bounded, non-empty

by Weierstrauss thm, the maximum of $g(z)$ on $\{z \mid \|z\|_2 \le 1\}$ must be finite.

Pf34: gradient of F(W,c) in least squares problem in neural network

Part 1: proving

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

In order to implement GD, we need the gradient of F.

Since $F : \mathbb{R}^{(d+1)p} \to \mathbb{R}$ is a function on the Euclidean space,

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

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

∂F

For $\frac{\partial F}{\partial w^{(j)}}$:

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

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

– Let $g_1(t) = (t - y_i)^2$, $g_2(w^{(j)}) = f_{W,c}(x^{(i)})$

Then $g_1 : \mathbb{R} \to \mathbb{R}$, $g_2 : \mathbb{R}^n \to \mathbb{R}$ and $(g_1 \circ g_2)(w^{(j)}) = g_1(g_2(w^{(j)}))$

$$= \left(f_{W,c}(x^{(i)}) - y_i\right)^2$$

By the chain rule:

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

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

So,

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

Pf34 Part 2: Proving:

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

- Since $f_{w,c}(x^{(i)}) = \langle c, \sigma(W^T x^{(i)}) \rangle$
$$= \sum_{k=1}^{p} c_k \sigma(\langle w^{(k)}, x^{(i)} \rangle)$$
$$= c_j \sigma(\langle w^{(j)}, x^{(i)} \rangle) + \sum_{\substack{k=1 \\ k \neq j}}^{p} c_k \sigma(\langle w^{(k)}, x^{(i)} \rangle)$$
((A

- Let $g_3(w^{(j)}) = \langle w^{(j)}, x^{(i)} \rangle$ $\quad (g_3 : \mathbb{R}^n \to \mathbb{R})$
$g_4(t) = c_j \sigma(t) + A$ $\quad (g_4 : \mathbb{R} \to \mathbb{R})$

Also, $(g_4 \circ g_3)(w^{(j)}) = g_4(g_3(w^{(j)})) = f_{w,c}(x^{(i)})$

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

Pf34. Part 3-

Altogether

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