Class Notes

Dustin Leatherman

April 26, 2020

Contents

1 Review & Introduction (2020/03/31)

1.1 Review

Orthogonal: Vectors are orthogonal when the dot product = 0.

1.1.1 Basis

$$\begin{aligned} \underset{(n\times 1)}{\vec{y}} &= \underset{(n\times p)}{A}\underset{(p\times 1)}{\vec{x}} \\ &= B\vec{c} \\ &= \Sigma c_i \vec{b_i} \text{ (most } c_i = 0) \end{aligned} \tag{1}$$

A: Basis Matrix
Properties of a Good Basis

- not all are orthogonal
- Allows for a sparse vector to be used ad the constant vector $\vec{c}$

Identity Matrices are the *worst* basis because most coefficients are non-zero.

2-Sparse Vector

$$\vec{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \\ 4 \end{bmatrix} \tag{2}$$

Very important!

> When dealing with Natural images and a good basis, there is a sparse vector.

1.1.2 Kernel

The kernel of a linear mapping is the set of vectors mapped to the 0 vector. The kernel is often referred to as the **null space**. Vectors should be linearly independent.

$$Ker(A) = \vec{x} \in \mathbb{R}^n \colon A\vec{x} = \vec{0} \tag{3}$$

A must be designed such that the Kernel of A does not contain any s-sparse vector other than $\vec{0}$

Main Idea: For (1), reduce $\vec{y}$ to a K-Sparse matrix to reduce the amount of non-zero numbers.

1.2 Linear Algebra Review

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \vec{v} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \tag{4}$$

$$\begin{aligned} \underset{(1\times3)(3\times1)}{\vec{u}^T\vec{v}} &= \begin{bmatrix} 1 & 2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = 1 + 2 - 2 = 1 \\ &= \vec{u} \cdot \vec{v} \end{aligned} \tag{5}$$

$$\underset{(3\times1)(1\times3)}{\vec{u}\ \vec{v}^T} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 4 \\ -1 & -1 & -2 \end{bmatrix} \tag{6}$$

$\vec{u}\ \vec{v}^T \neq \vec{u}^T\ \vec{v}$

1.2.1 Inner Product

$$\begin{aligned} <\vec{a}, \vec{b}> &= \vec{a} \cdot \vec{b} \\ &= \vec{a}^T \vec{b} \end{aligned} \tag{7}$$

1.2.2 Cauchy-Schwartz Inequality

$$\vec{a} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \tag{8}$$

$$\begin{aligned} | <\vec{a}, \vec{b}> | &\leq \sqrt{1^2 + 2^2 + (-1)^2} \times \sqrt{1^2 + 1^2 + 2^2} \\ | <\vec{a}, \vec{b}> | &\leq ||\vec{a}||_2\ ||\vec{b}||_2 \text{ (euclidean/l2-norm)} \end{aligned} \tag{9}$$

1.2.3 Norms

Why is the l1 norm preferred for ML opposed to the classic l2 norm?

Philosophically,

If we looked at a sphere in l2 norm, the shadow casted would be a circle regardless of the direction of the light.

Looking at a sphere in the l1 norm is shaped as a tetrahedron. The shadow cast by a tetrahedron is different for different angles so observing the shadow provides a lot more context about the sphere.

1. Euclidean/l2

   Sphere: $||\vec{x}||_2 = \sqrt{(-4)^2 + 3^2} = \sqrt{25} = 5$

   (a) FOIL Given 2 fixed vectors x,y. Consider the l2-norm squared:

   $$f(t) = ||x + ty||_2^2$$

$$
\begin{aligned}
f(t) =& ||x+ty||_2^2 \\
=& < x+ty, x+ty > \\
=& < x,x > +t < x,y > +t < y,x > +t^2 < y,y > \\
=& ||x||_2^2 + 2t < x,y > +t^2||y||_2^2
\end{aligned} \tag{10}
$$

Note: t<x,y> and t<y,x> can be combined because their dot-products are equivalent. $\vec{x} \cdot \vec{y} = \vec{y} \cdot \vec{x}$

When using Machine Learning, don't use l2 norms. Use l1

(b) Derivative

$$
\begin{aligned}
\frac{d}{dt}(||x+ty||_2^2) =& 2 < x,y > +2t||y||_2^2 \\
=& 2x^T y + 2ty^T y
\end{aligned} \tag{11}
$$

2. Simplex/l1

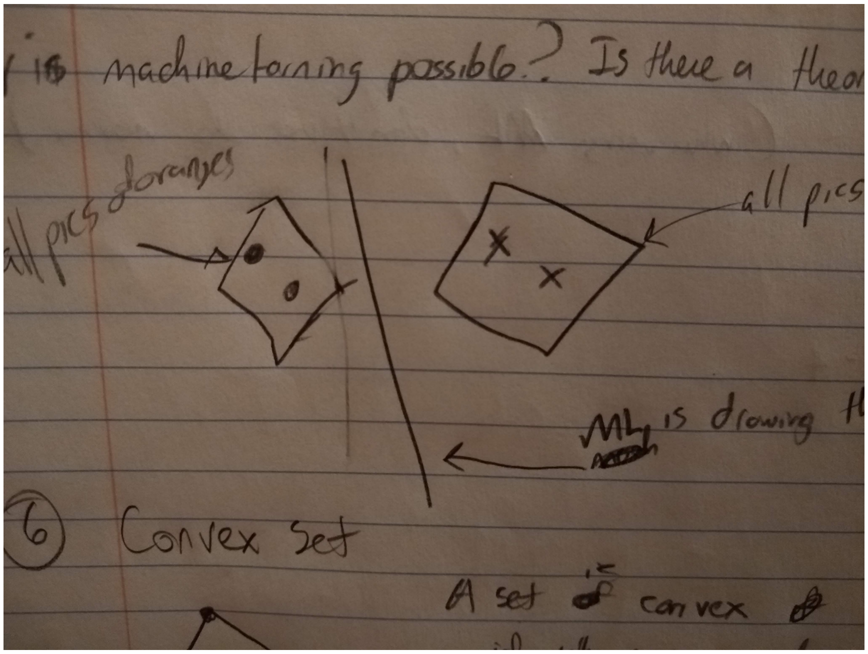
   Sphere: $||\vec{x}||_1 = |-4| + |3| = 7$

3. Infinity

   Sphere: $||\vec{x}||_\infty = Max|-4|, |3| = 4$

1.3 Optimization

Why is Machine Learning Possible? Is there a theoretical guarantee?



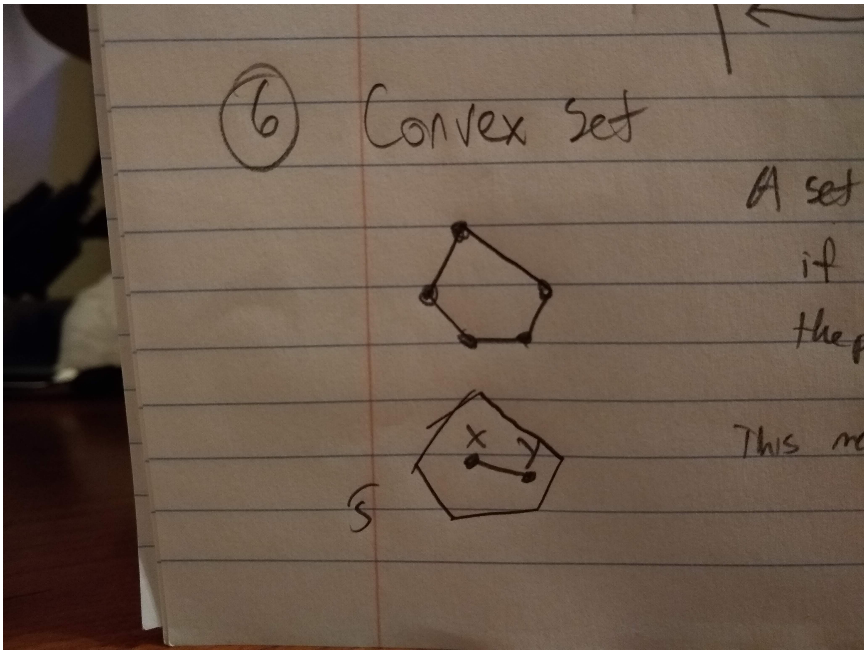
Imagine A is the set of all dogs and B is the set of all Cats

If the sets are convex and do not overlap, there exists a line between them which acts as a divider for determining whether a new pic belongs in A or B.

1.4 Convex Set

A set is convex if whenever X and Y are in the set, then for $0 \leq t \leq 1$ the points $(1-t)x + ty$ must also be in the set.

- #+ATTR$_{\LaTeX}$: scale=0.5



1.5 Separating Hyper-plane Theorem

Let C and D be 2 convex sets that do not intersect. i.e. the sets are **disjoint**.
Then there exists a vector $\vec{a} \neq 0$ and a number b such that.

$$a^T x \leq b \forall x \in C$$

and

$$a^T x \geq b \forall x \in D$$

The Separating Hyper-plane is defined as $x \colon a^T x = b$ for sets C, D.
This is the theoretical guarantee for ML

vector a is perpendicular to the plane b.

2 Why Separating Hyperplane Theorem & Subspace Segmentation Example (2020/04/07)

2.1 Why is Separating Hyper-plane Theorem true?

2.1.1 Math Background

Let $x = d - c,\ y = u - d$

1. Square of the \$l$_2$\$-norm is the inner product

$$\|x\|_2^2 = \langle x, x\rangle = x^T x$$

$$(d-c)^T(d-c) = \|d-c\|_2^2$$

2. Expansion of Vectors

$$\begin{aligned} &\|x+ty\|_2^2 \\ =&\langle x+ty, x+ty\rangle \\ =&\|x\|_2^2 + 2t\langle x, y\rangle + t^2\|y\|_2^2 \end{aligned} \tag{12}$$

3. Derivative of vector products

$$\frac{d}{dt}(\|x+ty\|_2^2) = 2x^T y + 2ty^T y$$

$$\frac{d}{dt}(\|x+ty\|_2^2)|_{t=0} = 2x^T y$$

$$\frac{d}{dt}(\|d+t(u-d)-c\|_2^2)|_{t=0} = 2(d-c)^T(u-d)$$

2.1.2 Separating Hyper-plane Theorem

C, D are convex disjoint sets. Thus there exists a vecto $\vec{a} \neq 0$ and a number b such that

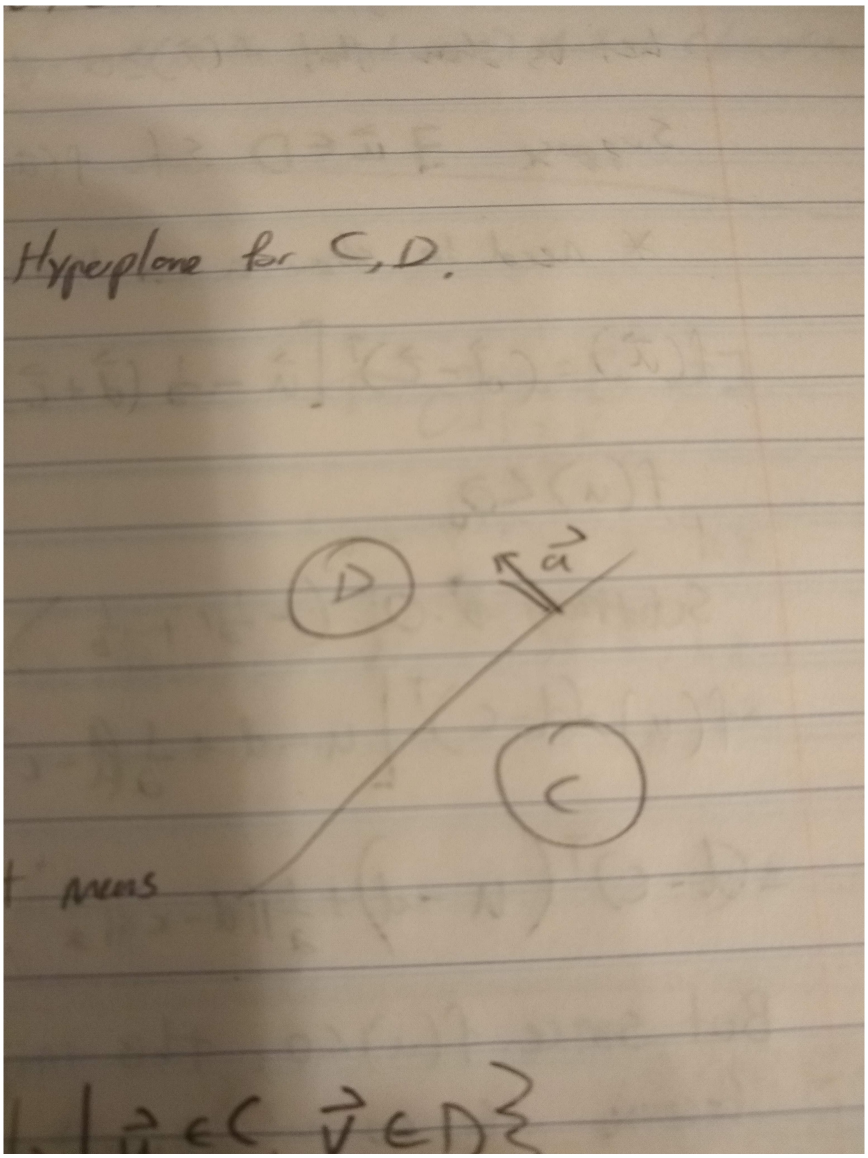
$$a^T x \leq b, \forall x \in C$$

and

$$a^T x \geq b, \forall x \in D$$

$x : a^T x = b$ is the separating hyper-plane for C,D.
When $b = 0$, then inconclusive answer.

2.1.3 Why is it true?



$$\begin{aligned} \vec{a}^T \vec{x} &\leq b \text{ on side C} \\ \vec{a^T} \vec{x} &\geq \text{ on side D} \end{aligned} \tag{13}$$

Goal: Prove $\vec{a}$ exists as that means a separating hyperplane exists.

$$dist(C, D) = min\|\vec{u} - \vec{v}\|_2 | \vec{u} \in C, \vec{v} \in D = \|\vec{c} - \vec{d}\|_2$$

where $\|\vec{u} - \vec{v}\|_2$ is the euclidean distance.
Let $\vec{a} = \vec{d} - \vec{c},\ b = \frac{1}{2}(\|\vec{d}\|_2^2 - \|\vec{c}\|_2^2)$
We will show that

$$f(\vec{x}) = a^T x - b$$

has the property that

$$f(\vec{x}) \leq 0,\ \forall \vec{x} \in C$$

and

$$f(\vec{x}) \geq 0,\ \forall \vec{x} \in D$$

Note: $(\vec{d} - \vec{c})^T \frac{1}{2}(\vec{d} + \vec{c}) = \frac{1}{2}(\|\vec{d}\|_2^2 - \|\vec{c}\|_2^2)$
What does showing something mean?
Let us show that $F(\vec{x}) \geq 0,\ \forall \vec{x} \in D$ (Argue by Contradiction)
Suppose $\exists \vec{u} \in D$ such that $f(\vec{x}) < 0$

$$f(\vec{u}) = (\vec{d} - \vec{c})^T [\vec{u} - \frac{1}{2}(\vec{d} + \vec{c})] = (\vec{d} - \vec{c})^T \vec{u} - \frac{1}{2}(\|\vec{d}\|_2^2 - \|\vec{c}\|_2^2)$$

Subtract 0

$$f(u) = (d - c)^T [u - d + \frac{1}{2}\|d - c\|]$$

$u - \frac{1}{2}d + \frac{1}{2}c$
$u - d + \frac{1}{2}d - \frac{1}{2}c$

$$f(u) = (d - c)^T (u - d) + \frac{1}{2}\|d - c\|_2^2$$

Now we observe that

$$\frac{d}{dt}(\|d + t(u - d) - c\|_2^2)|_{t=0} = 2(d - c)^T (u - d) < 0$$

and so for some small $t > 0$,

$$\|d + t(u - d) - c\|_2^2 < \|d - c\|_2^2$$

$g'(t) < 0$ means decreasing. Thus $g(t) < g(0)$.
Let's call point $p = d + t(u - d)$
Then

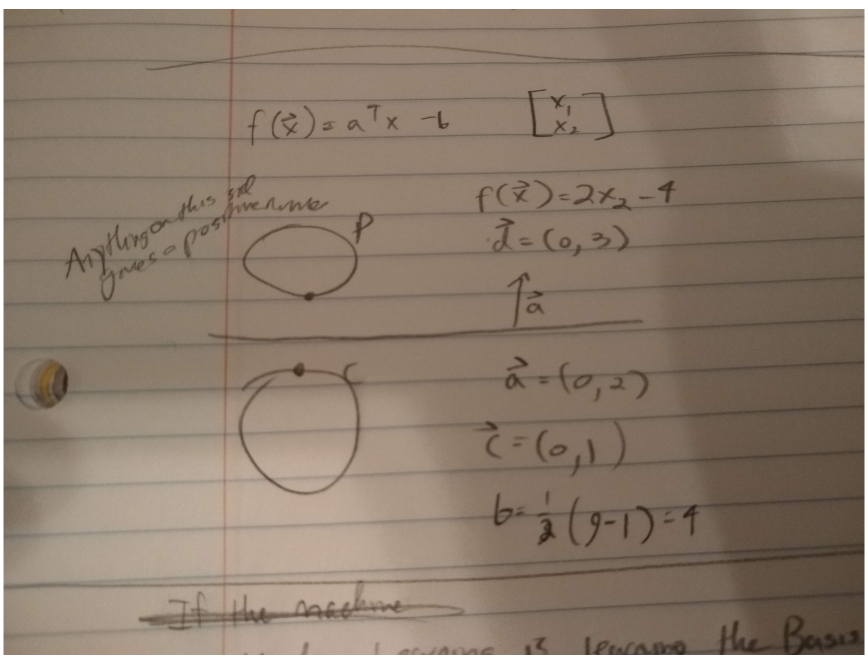
$$\|p - c\|_2^2 < \|d - c\|_2^2$$

This is a contradiction. Both d and u are in set D. Thus by the definition of convexity, $p = (1 - t)d + tu$

D is a convex set so p must also be in D. This situation is impossible since d is the point in D that is closest to c.

2.1.4 Example

Let $f(\vec{x}) = a^T x - b$



2.2 Subspace Segmentation Example

Machine Learning is learning the Basis A. If we can deduce that a vector $\vec{x}$ is a linear combination of A, then a vector is a subspace of Basis A and we know that it belongs to A.

$$V_1 = (x, y, z) \in R^3 : z = 0$$
$$V_2 = (x, y, z) \in R^3 : x = 0, y = 0$$

V_i is the affine variety (it is also a Ring, Module)

Apply a Veronase map with degree 2 to lift up from 3 to 6 dimensions.

$$\nu_n \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x^2 \\ y^2 \\ z^2 \\ xy \\ xz \\ yz \end{bmatrix}, \nu_n : R^3 \to R^6$$

$$\begin{aligned} z_1 = (3,4,0), z_2 = (4,3,0), \\ z_3 = (2,1,0), z_4 = (1,2,0), \\ z_5 = (0,0,1), z_6 = (0,0,3), z_7 = (0,0,4) \end{aligned} \tag{14}$$

Plug the sample points into the Veronase map to produce a matrix L

$$L = \begin{bmatrix} 9 & 16 & 4 & 1 & 0 & 0 & 0 \\ 16 & 9 & 1 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 9 & 6 \\ 12 & 12 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \in R^{6\times 7}$$

solve for $\vec{c}$, where $\vec{c}^T L = \vec{0}$

$$\vec{c_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \vec{c_2} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Rank(L) = 4 (since there are 4 linearly independent rows)

$$\begin{aligned} q_1(X) =& \vec{c}^T \nu_n(X) \\ =& xz \\ q_2(X) =& \vec{c}_2^T \nu_n(X) \\ =& yz \end{aligned} \tag{15}$$

We have:

$$
\begin{aligned}
q_1(X) &= xz \quad V_1 = (z = 0) \\
q_2(X) &= yz \quad V_2 = (x = 0, y = 0)
\end{aligned} \tag{16}
$$

Observe:
$V_1 \cup V_2 = ((x, y, z) \in R^3 : q_1(X) = 0, q_2(X) = 0)$
Construct the Jacobian matrix
J(Q)(X) $= \begin{bmatrix} \frac{\partial q_1}{\partial x} & \frac{\partial q_1}{\partial y} & \frac{\partial q_1}{\partial z} \\ \frac{\partial q_2}{\partial x} & \frac{\partial q_2}{\partial y} & \frac{\partial q_2}{\partial z} \end{bmatrix} = \begin{bmatrix} z & 0 & x \\ 0 & z & y \end{bmatrix}$

1. When $z = z_1 = (3, 4, 0), J(Q)(z_1) = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 4 \end{bmatrix}$

   When $z = z_3 = (2, 1, 0),\ J(Q)(z_3) = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 1 \end{bmatrix}$

   The right null space of $J(Q)(z_1)$ has basis $\vec{b}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \vec{b}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

2. When $z = z_5 = (0, 0, 1),\ J(Q)(z_5) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

   When $z = z_7 = (0, 0, 4),\ J(Q)(z_7) = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix}$ The right null space of $J(Q)(z_5)$ has basis $\vec{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

$C = [\vec{c}_1 | \vec{c}_2]$

3 Sparse Representation & Problem P0 . P1 (2020/04/14)

3.1 Big Idea

Your Data is a vector $x \in R^N$ where all vectors are column vectors. Each x is s-sparse i.e. each vector has at **most** s non-zero entries. Let s = 5000. We don't know where the non-zero entries are located.

Let $\underset{(m \times N)}{A}$, $m < N$
$N = 100,000,\ m = 20,000$
Short + Wide Matrix

This is the opposite of the kinds of matrices seen in Linear Regression which are tall and skinny.

What if we can design a matrix $A \in R^{m \times N}$ so that for each s-sparse $\vec{x} \in R^N$, you can store $\vec{y}$ instead? $(A\vec{x} = \vec{y})$

Q: Is there a way to get back $\vec{x}$ from $\vec{y}$? We observe $\vec{y}$.

A: Yes!

Properties of A

- A cannot be the 0 matrix.
- if $\vec{x}_1$ is s-sparse and $\vec{x} \neq 0$, what if $\vec{x}_1$ is in $ker(A)$? No! that would return $\vec{0}$ which means we cannot reconstruct the original matrix since there are multiple vectors in Ker(A).

Using Techniques from 1955

1. Is $\vec{x}$ the inverse of $\vec{y}$ or psuedo-inverse, or Moore-Penrose inverse, or...?

$$\begin{aligned} \vec{y} =& A\vec{x} \\ A^{\#}\vec{y} =& A^{\#}A\vec{x} \text{ where } \mathrm{A}^{\#}A = I \end{aligned} \tag{17}$$

Doesn't work! This is because there is no way to guarantee that $\vec{x}$ is a s-sparse vector.

1. Can we use gradient descent to solve for $\vec{x}$ to minimize $\|\vec{y} - A\vec{x}\|_2$

   No! Why?

   pick any vector $\vec{v} \in Ker(A)$. $\vec{y} = A(\vec{x} + \vec{v})$ however, $(\vec{x} + \vec{v})$ may not be sparse.

New math was needed to solve this problem so it was created in 2005 by Donoho, Candes, and Tao using the \$l$_1$\$-norm instead of the euclidean norm (l_2).

3.2 Background

\$l$_1$\$-norm: $\|x\|_1 = |x_1| + |x_2| + |x_3|$

\$l$_2$\$-norm: $\|x\| = \sqrt{|x_1|^2 + |x_2|^2 + |x_3|^2}$

For $\vec{x} \in R^n$, $\vec{y} \in R^N$, then

$$\|\vec{x}+\vec{y}\| \leq \|x\|_1 + \|y\|_1$$

For a norm to be valid, it must uphold the **Triangle Inequality**.
$\vec{a}$ is one side of a triangle, $\vec{b}$ is a second side, third side, ...

$$\begin{aligned} |\vec{a}+\vec{b}| &\leq |\vec{a}| + |\vec{b}| \\ \|\vec{x}+\vec{y}\|_1 &\leq \|\vec{x}\|_1 + \|\vec{y}\|_1 \\ \|\vec{x}+\vec{y}\|_2 &\leq \|\vec{x}\|_2 + \|\vec{y}\|_2 \\ \|\vec{x}+\vec{y}\|_2 &\leq \|\vec{x}\|_\infty + \|\vec{y}\|_\infty \end{aligned} \tag{18}$$

It also must be distributive:
If $\vec{x}_1 + \vec{x}_2 = \vec{y}$, then $(\vec{x}_1 + \vec{x}_2) \cdot \vec{a} = \vec{y} \cdot \vec{a}$ for any $\vec{a}$

$$\langle \vec{x}_1 + \vec{x}_2, \vec{a} \rangle = \langle \vec{y}, \vec{a} \rangle \to \langle \vec{x}_1, \vec{a} \rangle + \langle \vec{x}_2, \vec{a} \rangle = \langle \vec{y}, \vec{a} \rangle$$

3.3 Warm-up

$A = [\vec{a}_1|...|\vec{a}_N]$

$\|\vec{a}_j\|_2 = 1 = \langle \vec{a}_j, \vec{a}_j \rangle$

Let $\vec{v} \in Ker(A),\ \vec{v} \neq \vec{0},\ \vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ ... \\ v_N \end{bmatrix}$

Assume $\vec{a}_j$ are unit vectors.
Pick $i = 3$ observations.

1. Multiply by 1. Be Sneaky.

   $v_i = v_i \langle \vec{a}_i, \vec{a}_i \rangle$

2. $\vec{v} \in Ker(A)$

$$\begin{aligned} &v_1 a_1 + v_2 a_2 + ... + v_n a_n = \vec{0} \\ \to &\langle v_1 a_1 + ... + v_N a_N, a_i \rangle = \langle \vec{0}, a_i \rangle \\ \to &\langle v_1 a_1, a_i \rangle + ... + \langle v_N a_N, a_i \rangle = \langle \vec{0}, a_i \rangle \end{aligned} \tag{19}$$

Keep $v_3 \langle a_3, a_i \rangle$ on the left side. Move everything to the other side. Thus,

$$v_i = \langle v_i a_i, a_i \rangle = -\sum_{j=1, j \neq i} v_j \langle a_j, a_i \rangle$$

Since $i = 3$, $v_3 \langle a_3, a_i \rangle = v_i$

$$|v_i| \leq \sum_{j=1, ji} |v_j| \cdot |\langle a_j, a_i \rangle|$$

What is the absolute value of a single number in $Ker(A)$? There is a relation between v_i and the rest of the entries in $\vec{v}$.

Why "=" becomes $\leq$

For example, if -2 = 3 + (- 5), then

3.4 Getting Ready to Formulate the Problem

3.4.1 Problem P0

Find the s-sparse $\vec{x} \in R^N$ such that $\vec{y} = A\vec{x}$.

Ex. Problem 1 HW 1.

Find a 2-sparse vector $\vec{x} \in R^8$ such that $\vec{y} = A\vec{x}$.

There are $\binom{8}{2}$ 2-sparse vectors. (28).

Imagine N = 100,000 and s = 5000. Not feasible to try all sparse-vectors.

3.4.2 Problem P1 (Convex Optimization)

Given $A \in R^{m \times N}$ and measurement $\vec{y} = R^m$, solve the optimization problem,

$$\underset{x \in R^N}{min} \|x\|_1$$

subject to constraint $y = A\vec{x}$

Find a condition on matrix A, so that solving P1 will recover the s-sparse vector $x \in R^N$

3.5 Null Space Property of Order s

3.5.1 Setting up Notation

Let $\vec{v} \in Ker(A), \ \vec{v} \neq \vec{0}$

Let the set of indices , where $\vec{v}[j] \neq 0$ to be S.

e.g. $\vec{x} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 2 \\ 3 \\ 0 \\ 4 \end{bmatrix}$

$S = \{3, 5, 7\}$ (non-zero indices. Also called the support vector of $\vec{v}$).
$|S| = s$ (number of elements. i.e. sparsity)
$\bar{S} = \{1, 2, 4, 6\}$ (complement. i.e zero indices)

$$\vec{v} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ -2 \\ 2 \end{bmatrix}, \vec{v}_S = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 2 \\ 0 \\ 2 \end{bmatrix}, \ \vec{v}_{\bar{S}} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ -2 \\ 0 \end{bmatrix}$$

$\vec{v} = \vec{v}_S + \vec{v}_{\bar{S}}$

3.5.2 Definition

Let A be a $m \times N$ matrix.

Let S be a subset or $\{1, 2, 3, ..., N\}$. Suppose $N = 50$, and $S = \{3, 5, 7\}$

1. We say that a matrix A satisfies the null space property with respect to a set S if
$$\|\vec{v}_S\|_1 < \|\bar{S}\|, |\forall \vec{v} \vec{\in} Ker(A)$$

2. If it satisfies the null space property with respect to any set S of size s where S is a subset of $\{1, 2, 3, ..., N\}$. $s < N$

If a matrix satisfies this property, what does it buy us?

If a matrix A satisfies the Null Space property of order s, then solving problem P1 will solve P0. i.e. you can recover any s-sparse vector $\vec{x}$ from the measurement y where $\vec{y} = A\vec{x}$

> If A has a small coherence, then it satisfies the Null Space Property of order s.

Let $A = [\vec{a}_1|...|\vec{a}_N]$

$$\mu_1 = \underset{j \neq k}{max} |\langle \vec{a}_j, \vec{a}_k \rangle|$$

Assume $\vec{a}_j$ has l_2-norm equal to 1.

3.5.3 Theorem

Same assumptions as above.

Suppose $\mu_1 \cdot s + \mu_1 \cdot (s-1) < 1$

The matrix satisfies the Null Space property of order s.

Remarks

1. $\mu_1(2s-1) < 1$ if true, then A satisfies NSP of order s. It is not a necessary condition. It is a sufficient condition.

2. From the warm up, if we fix an index i, then for $\vec{v} \in Ker(A)$,

$$|v_i| \leq \sum_{j=1, j \neq i} |v_j| \cdot |\langle \vec{a}_j, \vec{a}_i \rangle| \tag{20}$$

1. Note that $|v_i|$ is just one term in $\|v\|_1$ because

$$\|v\|_1 = |v_1| + |v_2| + ...$$

3.5.4 Proof

Given A is an $m \times N$ matrix. $A = [\vec{a}_1|...|\vec{a}_N]$.

Suppose $\|\vec{a}_j\| = 1,\ \mu_1 \cdot s + \mu_1 \cdot (s-1) < 1$

Show that NSP of order s holds.

i.e.

$$\|\vec{v}_S\| < \|\vec{v}_{\bar{S}}\|, \forall \vec{v} \in ker(A)|\{\vec{0}\}$$

and for every set

$$S \subset \{1, 2, 3, ..., N\} \text{with} |S| = s$$

Let $\vec{v} = Ker(A)$

$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ ... \\ v_n \end{bmatrix}$

$A\vec{v} = v_1\vec{a}_1 + ... + v_N\vec{a}_N = \vec{0}$
Let $S \subset \{1, 2, ..., N\}, \ |S| = s$. Pick any $\vec{a}_i, i \in S$
Then $v_i = v_i\langle\vec{a}_i, \vec{a}_i\rangle$. Also, $v_1\langle\vec{a}_i, \vec{a}_i\rangle + ... + v_N\langle\vec{a}_N, \vec{a}_i\rangle = 0$

$$\begin{aligned}
\rightarrow v_i = v_i\langle\vec{a}_i, \vec{a}_i\rangle = -\sum_{j=1, j\neq i} v_i\langle\vec{a}_j, \vec{a}_i\rangle \\
\rightarrow v_i = -\sum_{l\in S} v_l\langle\vec{a}_l, \vec{a}_i\rangle - \sum_{j\in S, j\neq i} v_j\langle\vec{a}_j, \vec{a}_i\rangle \\
\rightarrow |v_i| \leq \sum_{l\in S} |v_l||\langle\vec{a}_l, \vec{a}_i\rangle| + \sum_{j\in S, j\neq i} |v_j||\langle\vec{a}_j, \vec{a}_i\rangle|
\end{aligned} \tag{21}$$

sum over all $i \in S$ to get
$\|\vec{v}_S\|_1 = \sum_{i\in S} |v_i|$

This adds up all the inequalities for one inequality to rule them all.

$$\begin{aligned}
&\leq \sum_{i\in S}\sum_{l\in \bar{S}} |v_l| \cdot |\langle\vec{a}_l, \vec{a}_i\rangle| + \sum_{i\in S}\sum_{j\in S, j\neq i} |v_j| \cdot |\langle\vec{a}_j, \vec{a}_i\rangle| \\
&= \sum_{l\in \bar{S}} |v_l| \sum_{i\in S} |\langle\vec{a}_l, \vec{a}_i\rangle| + \sum_{j\in S} |v_j| \sum_{i\in S, i\neq j} |\langle\vec{a}_j, \vec{a}_i\rangle| \\
&\leq \sum_{l\in S} |v_l|\mu_1 \cdot s + \sum_{j\in S} |v_j|\mu_1(s-1) \\
\|\vec{v}_S\|_1 &\leq \mu_1 \cdot s\|\vec{v}_{\bar{S}}\| + \mu_1(s-1)\|\vec{v}_{\S}\|
\end{aligned} \tag{22}$$

$$(1 - \mu_1(s-1))\|\vec{v}_{\bar{S}}| \ \leq \mu_1 \cdot s\|\vec{v}_S\|$$

Since $\mu_1(s-1) + \mu_1(s) < 1$ by hypothesis, so $1 - \mu_1(s-1) \geq \mu_1(s)$ and hence $\|\vec{v}_S\|_1 < \|\vec{v}_{\bar{S}}\|_1$

3.6 Ways to Solve P1

There are 8 algos to solve P1. The worst performing one is Linear programming.

This is one of the Algos

3.6.1 Algos

$A = \begin{bmatrix} 1 & 1 \end{bmatrix}$ $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$a_{11} = a_{12} = 1$

$Q = \begin{bmatrix} \frac{1}{w_1} & 1 \\ 0 & \frac{1}{w_2} \end{bmatrix}$

1. Minimize $\|\vec{x}_1\|$ subject to $\vec{y} = A\vec{x}$

$$\begin{aligned} \vec{y} =& (AA^T)(AA^T)^{-1}\vec{y} \\ \vec{y} =& A(A^T(AA^T)^{-1}\vec{y}) \end{aligned} \tag{23}$$

Why not let $\vec{x} = (A^T(AA^T)^{-1}\vec{y})$
maybe we can do better.
$\vec{y} = AQA^T(AQA^T)\vec{y}$
Why not let $\vec{x} = (QA^T(AQA^T)^{-1}\vec{y})$
How to choose Q?

1. $min \sum_{i=1}^{N} W_i x_i^2$ subject to $\vec{y} = A\vec{x}$

   This is not the \$l₁\$-norm but it would be if $w_i = \frac{1}{|x_i|}$.

   solve 2. then substitute w_i

2. min: $w_1 x_1^2 + w_2 + x_2^2$ subject to $y = a_{11}x_1 + a_{12}x_2$

   $f(x_1) = w_1 x_1^2 + w_2(y - x_1)^2$

   $f'(x_1) = 0$ solve for x_1

   $2w_1 x_1 + 2(y - x_1)(-1)w_2 = 0$

   $x_1 = \frac{w_2}{w_1+w_2} y, \; x_2 = \frac{w_1}{w_1+w_2} v$

$$\begin{aligned} AQA^T =& \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{w_1} & 0 \\ 0 & \frac{1}{w_2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ =& \frac{w_1 + w_2}{w_1 w_2} \end{aligned} \tag{24}$$

$$QA^T(AQA^T)^{-1}y = \begin{bmatrix} \frac{1}{w_1} \\ \frac{1}{w_2} \end{bmatrix} \frac{w_1 w_2}{w_1 + w_2} y \tag{25}$$

4 Sparse Representation pt 2 (2020/04/21)

4.1 Historical Perspective

Why is the visual system so powerful? Hypothesis is our brain uses sparse representation of Visual Data.

Let a picture $\vec{y} = c_1\vec{b}_1 + ... + c_n\vec{b}_n$
so that most c_j are zero.
Sparse representation used to be called Sparse Coding.
Robust Facial Recognition uses Sparse Subspace Clustering.
Given 19 x 19 images, let $Y = [\vec{Y}_1|...|\vec{Y}_{45}],\ \vec{y}_j \in R^{361}$
19 * 19 = 361
Given Y, solve for matrix C

$$Y = YC,\ diag(C) = \vec{0}$$

Since we don't want $Y_i = Y_i$, that is why the constraint $diag(C) = \vec{0}$ is introduced. It ensures that a group of vectors can be a linear combination of others.

Each column of C is sparse since we want all column vectors to be a linear combination of a smaller set of columns.

4.2 Example - Handwritten Digit Recognition

Given 28 x 28 images, Let $B = [\vec{y}_1|...|\vec{y}_{4000}]$ where each $\vec{y}_j \in R^{784}$

- 800 images of 0, 1-800
- 800 images of 1, 801-1600
- 800 images of 2, 1601-2400
- 800 images of 3, 2401-3200
- 800 images of 8, 3201-4000

Let $\vec{f}$ be a new image of 2. Solve for X such that $\vec{f} = B\vec{x}$
Assume $\vec{x}$ is 20-sparse.
We would like to see the only **non-zero** entries at position 1601-2400.
Columns outside the range may be non-zero as well. There is a 95% probability that a digit will be 2, 5% it will be another digit.

4.2.1 Qualitative Theorem

Given $A^{m\times N}$ with $m << N$. If A is a Gaussian random matrix, then with overwhelming high probability, it satisfies some Exact Recovery Condition for s-sparse Vectors.

For most large undetermined systems of linear equations, the minimal l_1-norm solution is also the sparsest solution.

Topics of Research:

- Theory of Random Matrices
- Banach Spaces

4.3 Solving P1 solves P0. Why?

PO

Find the s-sparse $\vec{x} \in R^N$ such that $\vec{y} = A\vec{x}$.

P1

$A \in R^{m\times N}$ and measurement $\vec{y} \in R^m$. Solve optimization problem,

$$\min_{x\in R^N} \|x\|_1$$

subject to the constraint $y = A\vec{x}$

Suppose $\vec{y} = A\vec{x}$ and $\vec{y} = A\vec{z}$. Suppose $\vec{x}$ is a sparse vector and $\vec{z}$ is **not**.

We want to show that $\|\vec{x}\|_1 < \|\vec{x}\|_1$ - Null Space property of order S

$\|\vec{x}\|_1 = \|\vec{x} - \vec{z}_S + \vec{z}_S\|_1$ - $\vec{z}$ restricted to some Set S. (Subtract 0 so we can use triangle inequality).

Let $\vec{v} = \vec{x} - \vec{z},\ \vec{v} \in Ker(A)$

$A(\vec{x} + \vec{z}) = A\vec{v} = \vec{0}$

$$\begin{aligned}
\|\vec{x}\|_1 \leq & \|\vec{x} - \vec{z}_S\|_1 + \|z_S\|_1 && \text{(26a)}\\
= & \|\vec{v}_S\|_1 + \|\vec{z}_S\|_1 && \text{(26b)}\\
< & \|\vec{z}_S\|_1 + \|\vec{v}_{\bar{S}}\|_1 \qquad \text{via Null Space Property} && \text{(26c)}\\
= & \|-\vec{z}_{\bar{S}}\|_1 + \|z_S\|_1 \qquad \|x_{\bar{s}}\|_1 = 0 \text{ since x is sparse} && \text{(26d)}\\
= & \|\vec{z}\|_1 && \text{(26e)}
\end{aligned}$$

4.4 Adjoint

Let $T\colon V \to W$. For example, T can be a matrix from R^3 to R^2. In this case, V is R^3 and W is R^2

We write T^* for the adjoint of T.

$$\forall x \in V,\ \forall y \in W,\ \langle Tx, y\rangle = \langle x, T^* y\rangle$$

Horrible way to think of it, when T is a matrix, the adjoint is the same as the transpose.

Q: When A is an orthogonal matrix, what is A^*A? I

Hint: each column has \$l$_2$\$-norm 1, distinct cols are perpendicular.

Q: When A is an orthogonal matrix, why is $\|Ax\|_2 = \|x\|_2$ for every vector x? (This is known as an isometry)

$$\|Ax\|_2^2 = \langle Ax, Ax\rangle = \langle x, A^*Ax\rangle = \langle x, x\rangle = \|x\|_2^2$$

4.5 Restricted Isometry Property (RIP)

$A \in R^{m\times N}$ satisfies the restricted isometry property of order s and level δ_s $(0 < \delta_s \leq 1)$

$$(1-\delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_s)\|x\|_2^2,\ \forall \text{ s-sparse } x \in R^N$$

> Any s columns of the matrix A are **nearly** orthogonal to each other.

Q: What can we say about $|\langle (I - A^*A)x, x\rangle|$ when vector is s-sparse?

This is a small number.

Let $u,\ v \in R^N$ and $S \in \{1, 2, 3, ..., N\},\ |S| = s$

What can we say about the following?

$$|\langle u, (I - A * A)v\rangle|$$

We would like to be able to say

$|\langle u, (I - A^*A)v\rangle| \leq \delta_t \|u\|_2 \|v\|_2$

4.5.1 How to think about RIP?

Suppose A satisfies the restricted isometry property of order s.

Intuition: **Hopefully**, the matrix A^*A behaves like the Identity Matrix. \$(I

- A*A)$ is small.

If you take some s-sparse vector $\vec{x}$ and multiply it by $I - A^*A$, hopefully, the resulting vector will also be small.

4.5.2 Algorithm

Consider the following vectors,

$$\vec{x}_1 = \begin{bmatrix} 10 \\ -20 \\ 3 \\ -4 \\ 5 \\ -6 \\ -7 \\ 8 \\ 4 \end{bmatrix}, \ \vec{x}_2 = \begin{bmatrix} 10 \\ -20 \\ 0 \\ 0 \\ 0 \\ 0 \\ -7 \\ 8 \\ 0 \end{bmatrix}$$

<u>Hard Threshold</u>

$\tau_s(\vec{x})$ is the vector that keeps the s entries that are the largest in Absolute Value.

Example: When $s = 4, \ \tau_s(\vec{x}_1) = \vec{x}_2$

$\tau_s(\cdot)$ is an operator that takes a vector and will output a sparse vector.

$$\begin{aligned} \vec{u}_n =& \vec{x}_n + A^*(\vec{y} - A\vec{x}_n), \text{ where } \vec{y} = A\vec{x} && \text{(27a)} \\ =& \vec{x}_n + (A^*A\vec{x} - A^*A\vec{x}_n) && \text{(27b)} \\ =& (I - A^*A)\vec{x}_n + A^*A\vec{x} && \text{(27c)} \end{aligned}$$

- expect $\vec{u}_n$ close to $\vec{x}$
- however, $\vec{u}_n$ may not be sparse. Thus use $\tau_s(\cdot)$

  <u>Iterative Hard Thresholding</u>

$$\vec{x}_{n+1} = \tau_x(\vec{x}_n + A^*(\vec{y} - A\vec{x}_n))$$

4.6 Operator Norm

$\|A\| = \underset{x \neq 0}{max} \frac{\|Ax\|_2}{\|x\|_2}$

How much influence does A have on a vector x? Shrink, stretch, compress?

Describes how big a matrix is. If A is 2 x 3, then take $\vec{x} \in R^3,\ x \neq 0$

What is

$$\|A\| = max\{\|Ax\|_2 \colon \|x\|_2 = 1\}$$

4.6.1 Inner Product

Let A be a matrix . The inner product of two vectors Ax and y has this property,

$$|\langle Ax, y\rangle| \leq \|A\| \cdot \|x\|_2 \|y\|_2$$

Where $\|A\|$ is the operator norm of A.

By Cauchy-Schwartz Inequality,

$$\|\langle Ax, y\rangle\| \leq \|Ax\|_2 \cdot \|y\|$$

By def,

$$\|Ax\| \leq \|A\| \cdot \|x\|_2$$

Thus,

$$\|\langle Ax, y\rangle\| \leq \|A\| \cdot \|x\|_2 \cdot \|y\|_2$$