Lab 10

DSGA-1014: Linear Algebra and Optimization

CDS at NYU
Zahra Kadkhodaie

Fall 2021

A with linearly independent columns

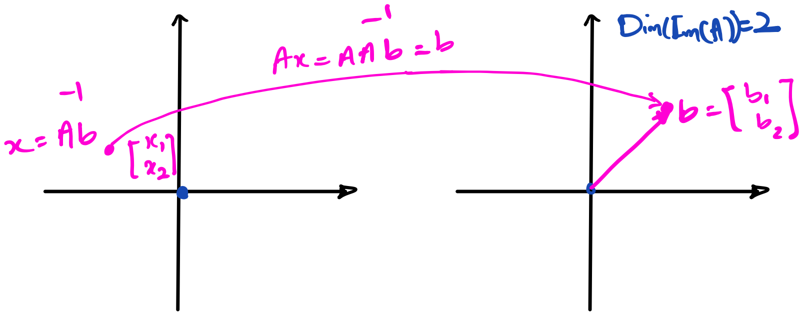
Assume $A \in R^{n \times n}$ is full rank. We have learned that $Ax = b$ has a unique solution.

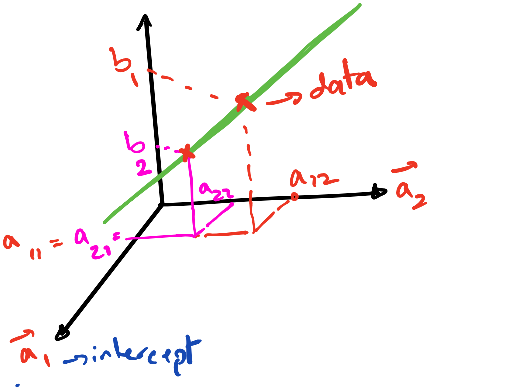
Example : $A \in R^{2 \times 2}$

Regression :

Find the best fitting line

Column view :


$x = A^{-1}b$
$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
$Ax = AA^{-1}b = b$
$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$
Dim(Im(A))=2

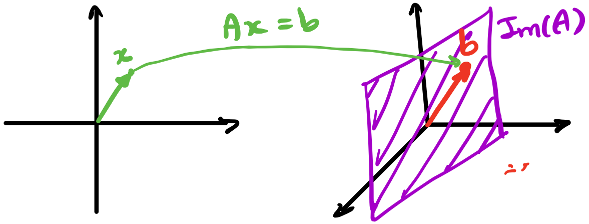


$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

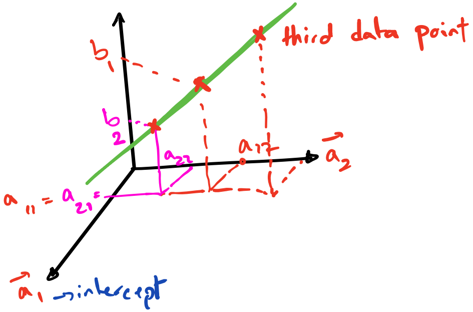
A with linearly independent columns

Assume $A \in R^{n \times m}$ is full rank, where $n > m$ (i.e. A is a tall matrix). In this case $Ax = b$ is a system of equations with too many rows (i.e. more equations than variables). We call this system of equations over-determined, which happens a lot in practice. Describe the solution of this system.

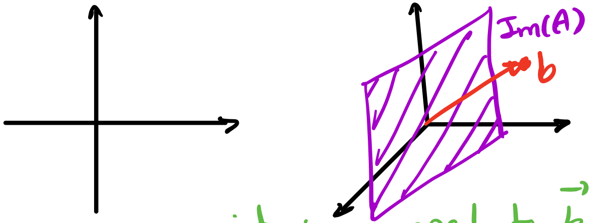
Case 1) Unique solution: Example: $A \in R^{3 \times 2}$



Unlikely case.
There is always noise in measurments

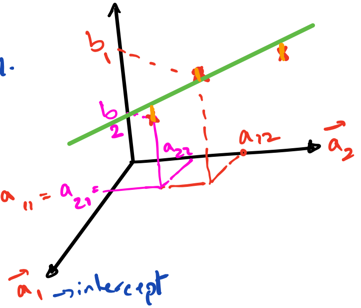


Case 2) No solution
$A \in R^{3 \times 2}$



No point is mapped to $\vec{b}$

This is what we normally deal with.

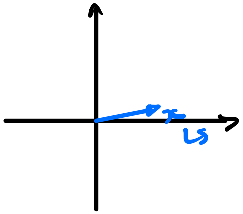


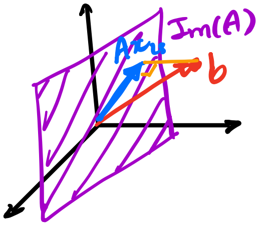
A with linearly independent columns

How do we solve $Ax = b$ when $b \notin Im(A)$? We can't! So we compromise and find the next best vector: x_{LS} such that Euclidean distance between Ax_{LS} and b is minimum. That is, x_{LS} is mapped to a vector on $Im(A)$ which is as close to b as possible.

$$x_{LS} = \text{argmin}_x ||Ax - b||^2$$

Show that Ax_{LS} is the projection of b onto $Im(A)$.
Show that error, $e = Ax - b$, is in $Im(A)^{\perp}$.





$f(x) = ||Ax - b||^2$ is a convex function

So we can find the min analytically:

$$x_{LS} = (A^TA)^{-1}A^Tb \quad \text{(from lecture 10)}$$

$$\Rightarrow Ax_{LS} = \underbrace{A(A^TA)^{-1}A^T}_{\text{Projection matrix onto } Im(A) \text{ (from lab 8)}} b = Pb$$

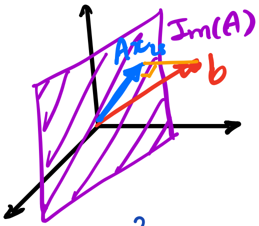
Take $z \in R^m$ s.t. $Az \in Im(A)$

$$\langle Az, e \rangle = (Az)^T e = z^T A^T (Ax_{LS} - b) = z^T (A^T A x_{LS} - A^T b)$$

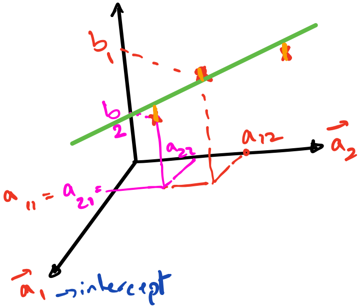
$$= z^T (\underbrace{A^T A (A^T A)^{-1}}_{I} \underbrace{A^T b - A^T b}_{0}) = 0$$

$\Rightarrow e$ is perpendicular to Az for $\forall z \in R^m$

$\Rightarrow e \in Im(A)^{\perp}$



$$\|e\|^2 = \|b\|^2 - \|Ax_{LS}\|^2$$

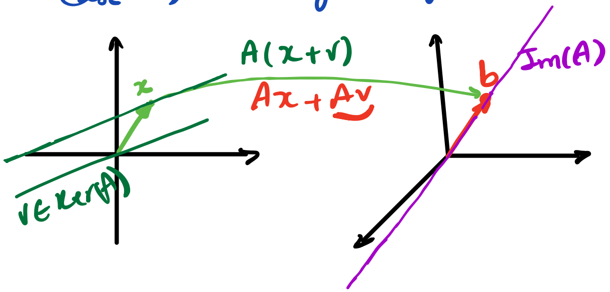


$$\|e\|^2 = e_1^2 + e_2^2 + e_3^2$$

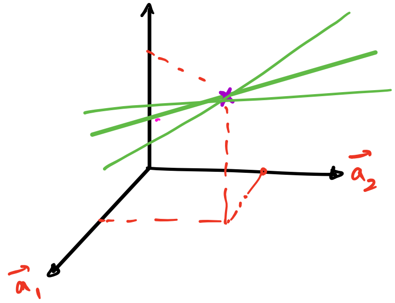
A with linearly dependent columns

The least square solution presented above only works if $A^T A$ is invertible. Since the $Ker(A^T A) = Ker(A)$, the least square as defined above only exists when columns of A are independent. Describe possible solutions.

Case 1) Infinitely many solutions. Example: $A \in R^{3x2}$ rank(A) = 1

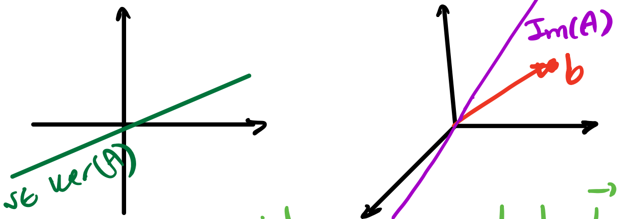


many lines satisfy the solution



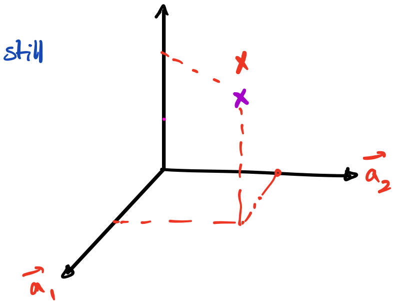
Case 2) No solution

$A \in R^{3x2}$



No point is mapped to $\vec{b}$

Note: The equation $A^TA x = A^Tb$ still must have a solution because $A^Tb \in Im(A^T) = Im(A^TA)$ So there must be an x that minimizes the objective



A with linearly dependent columns: pseudo-inverse

To solve this problem, we define pseudo-inverse as $A^\dagger = V\Sigma' U^T$ where $\Sigma' \in R^{d\times n}$ with $\Sigma'_{ii} = 1/\Sigma_{ii}$ if $\Sigma_{ii} \neq 0$, and zero otherwise. Show that $A^\dagger \in R^{d\times n}$ is the only matrix in $R^{d\times n}$such that

1. $AA^\dagger A = A$
2. $A^\dagger AA^\dagger = A^\dagger$
3. $AA^\dagger \in R^{n\times n}$ and $A^\dagger A \in R^{d\times d}$ are symmetric matrices.

$$\underset{n\times d}{\Sigma} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \; 0 \\ & 0 & \end{bmatrix} \Big\}\, n-d \qquad \Sigma' = \begin{bmatrix} 1/\sigma_1 & \cdots & 1/\sigma_r \; 0 & \end{bmatrix}$$

Note: If A invertible: $A^+ = A^{-1}$

$A = U\Sigma V^T \qquad A^+ = V\Sigma' U^T$

1) $AA^+A = U\Sigma \underbrace{V^T V}_{I} \Sigma' \underbrace{U^T U}_{I} \Sigma V^T = U \underbrace{\underset{n\times d}{\Sigma}\, \underset{d\times n}{\Sigma'}}\, \underset{n\times d}{\Sigma} V^T = U\Sigma V^T = A$

$$r\text{th}\rightarrow \underset{n\times n}{\begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \\ & & & & & \ddots \end{bmatrix}} \underset{n\times d}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \; 0 \end{bmatrix}} = \Sigma_{n\times d}$$

2) $A^{+}AA^{+} = V\Sigma' U^T U\Sigma V^T V\Sigma' U^T = V\Sigma'\underbrace{\Sigma\Sigma'}_{n\times n}U^T = V\Sigma' U^T = A^{+}$

3) $A^{+}A = V\Sigma' U^T U\Sigma V^T = V\underbrace{\Sigma'\Sigma}_{d\times d}V^T = \underbrace{\left[\begin{array}{ccccc} | & & | & & | \\ \overbrace{v_1 & \cdots & v_r}^{\text{basis for row space}} & \cdots & v_d \\ | & & | & & | \end{array}\right]\begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}\begin{bmatrix} | & & | & \\ v_1 & \cdots & v_r & \cdots \\ | & & | & \end{bmatrix}^T}_{\text{projection onto rowspace}(A)}$

$$\begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}$$

$AA^{+} = U\Sigma V^T V\Sigma' U^T = U\underbrace{\Sigma\Sigma'}_{n\times n}U^T$ projection onto columnspace of A

A with linearly dependent columns: pseudo-inverse

Using pseudo-inverse, we define the least square solution as $x_{LS} = A^{\dagger} y$.

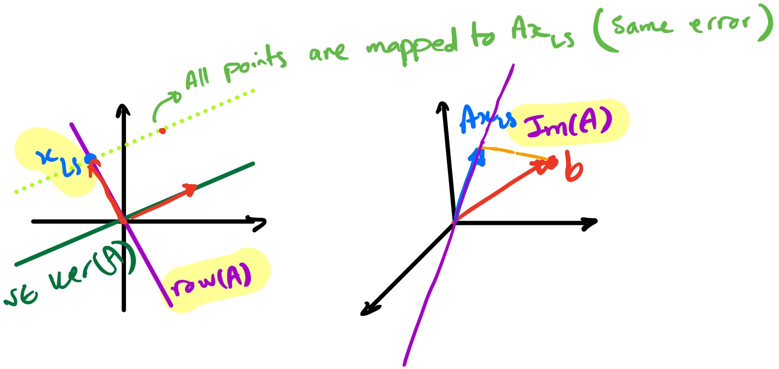
1. Show that when columns of A are independent the two least square solutions are the same.
2. Show that x_{LS} is always in the row space of A.
3. Give the set of all vectors that minimize $||Ax - y||^2$?

1) Show : $(A^T A)^{-1} A^T y = A^+ y$

$$(V \Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma U^T = V (\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T$$

$$\begin{bmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_d^2 \end{bmatrix}$$

$$= V \underbrace{(\Sigma^T \Sigma)^{-1}}_{d\times d} \underbrace{\Sigma^T}_{d\times n} U^T = \overbrace{V \begin{bmatrix} 1/\sigma_1 & & & \\ & \ddots & & 0 \\ & & 1/\sigma_d & \end{bmatrix} U^T}^{A^+}$$



If Colums of A are independent

$\Rightarrow$ rank(A) = m $\Rightarrow$ row space = domain

$\Rightarrow$ every x including $x_{LS} \in$ row space(A)

If Columns of A are not independent : $x_{LS} = A^{+} y = V \Sigma' U^T y$

$y' \in R^m$

$$= \begin{bmatrix} \underbrace{\begin{matrix} | & | & & | \\ v_1 & v_2 & \cdots & v_r \\ | & | & & | \end{matrix}}_{\text{basis for row(A)}} & \underbrace{\begin{matrix} & | \\ \cdots & v_m \\ & | \end{matrix}}_{\substack{m-r \\ \text{basis for Ker(A)}}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1}\langle u_1, y\rangle \\ \frac{1}{\sigma_2}\langle u_2, y\rangle \\ \vdots \\ \frac{1}{\sigma_r}\langle u_r, y\rangle \\ 0 \\ 0 \\ \vdots \end{bmatrix}_{m\times 1} \left.\vphantom{\begin{matrix}0\\0\\ \vdots\end{matrix}}\right\} m-r = \sum_{i=1}^{r} \frac{1}{\sigma_i}\langle u_i, y\rangle \vec{v}_i$$

$\forall z \in Ker(A)$: $A(x_{LS} + z) = Ax_{LS} + Az = Ax_{LS}$

All points in affine space $Ker(A) + x_{LS}$ give the same error.

A with linearly dependent columns: pseudo-inverse

Note: pseudo-inverse is particularly useful when $A \in R^{n \times m}$ is a short matrix ($n < m$). In this case, $Ax = b$ is an under-determined system of equations and even if A is full rank, rank(A) = n, columns of A are not independent and $A^T A$ is not invertible.

Ridge regression

Sometimes the objective deviates from least square solution. In Ridge regression, we add a penalty term to least square objective to promote a solution with small norm.

$$x_{ridge} = \text{arg min }_x ||Ax - b||^2 + \lambda ||x||^2$$

Show that x_{ridge} is in the row space of A.

$$x_{ridge} = (A^T A + \lambda I_m)^{-1} A^T y$$

$$= (V\Sigma^T U^T U \Sigma V^T + V \Lambda V^T)^{-1} V \Sigma^T U^T y$$

$$= (V\Sigma^T\Sigma V^T + V \Lambda V^T)^{-1} V \Sigma^T U^T y$$

$$= (V(\Sigma^T\Sigma + \Lambda) V^T)^{-1} V \Sigma^T U^T y$$

$$= V(\Sigma^T\Sigma + \Lambda)^{-1} V^T V \Sigma^T U^T y$$

$$= V(\underbrace{\Sigma^T\Sigma}_{m\times m} + \underset{m\times m}{\Lambda})^{-1} \Sigma^T U^T y$$

$$\underbrace{\begin{array}{c} \\ \\ \\ m-r\{ \end{array}\begin{bmatrix} \sigma_1^2+\lambda & & & \\ & \ddots & & \\ & & \sigma_r^2+\lambda & \\ & & & \lambda \ddots \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \ddots \end{bmatrix}}$$

$$\begin{array}{c} \\ \\ m-r\{ \end{array}\begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda} & & \\ & \ddots & \\ & & \frac{\sigma_r}{\sigma_r^2+\lambda} \\ & & & 0 \ddots \end{bmatrix}_{m\times n}.$$

$$x_{ridge} = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2+\lambda} \langle u_i, y \rangle \vec{v_i}$$

↳ $v_1, \ldots, v_r$ (basis for row space)

Ridge regression

Ax_{ridge} is no longer an orthogonal projection of b onto the Im(A). It is a modified projection where the component of the data in the direction of each left singular vector of the feature matrix is shrunk by a factor of $\sigma_i^2/(\sigma_i^2+\lambda)$ where σ_i is the corresponding singular value. Show that

$$Ax_{ridge} = \sum_{i=1}^{m} \frac{\sigma_i^2}{\sigma_i^2+\lambda}\langle b, u_i\rangle u_i$$

where u_i are the left singular vectors of A.

From previous question:

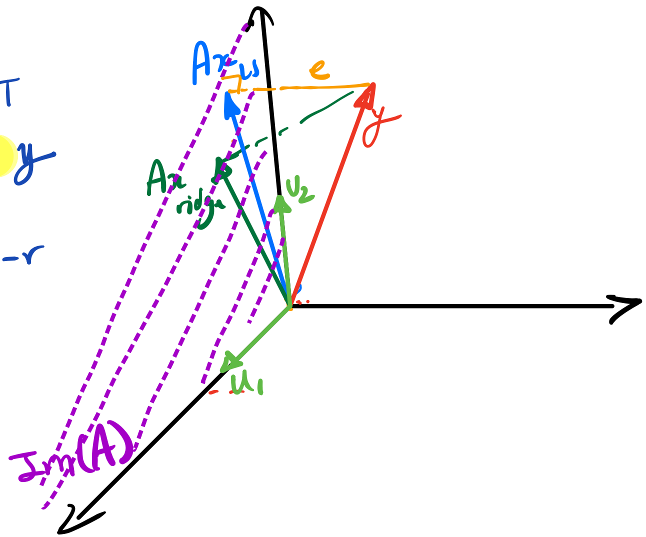
$$x_{ridge} = V(\underbrace{\Sigma^T\Sigma}_{m\times m} + \underset{m\times m}{\Lambda})^{-1}\Sigma^T U^T y$$

$$\Rightarrow Ax_{ridge} = U\Sigma V^T V(\underbrace{\Sigma^T\Sigma}_{m\times m} + \underset{m\times m}{\Lambda})^{-1}\Sigma^T U^T y$$

$$= U \underbrace{\Sigma (\Sigma^T \Sigma + \lambda)^{-1} \Sigma^T} U^T y$$

$$= U \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & & & & \\ & \ddots & & & & \\ & & \frac{\sigma_r^2}{\sigma_r^2+\lambda} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & \end{bmatrix}_{n \times n} \Big\} n-r \quad U^T y$$

$$= \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2+\lambda} \langle v_i , y \rangle \vec{u}_i$$



This reduces the influence of the directions corresponding to smaller singular values which are the ones responsible for more noise amplification.