Lecture 8 Least Squares

Note by Samion Suwito on 2/13/25

Last Time

$A = U_r \Sigma_r V_r^\top$, $A^+ = V_r \Sigma_r^{-1} U_r^\top$
Remark
$A^+ A = V_r V_r^\top =$ Projection onto $\mathcal{N}(A)^\perp$
$AA^\top = U_r U_r^\top =$ Projection onto $\mathcal{R}(A)$
(identity minus the above is the projection onto the subspace perp)
Lecture 7 Low Rank Approximation and Pseudo inverse > Examples

Connect to Least Squares

For $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ find x^* of minimum l^2 norm s.t. It minimises:

$$\min_x \|Ax - b\|_2^2$$

Important Idea: If A doesn't have full column rank then there is a non trivial null space meaning that the solutions won't be unique as you can add a vector from the null space to x.

$$\begin{aligned} \min_x \|Ax - b\|_2^2 &= \min_{y \in \mathcal{R}(A)} \|y - b\|_2^2 \\ &= \|\pi_{\mathcal{R}(\mathcal{A})}(b) - b\|_2^2 \\ &= \|AA^+ b - b\|_2^2 \end{aligned}$$

Projection is AA^+ shown above.

Solutions to the problem are those x such that $Ax = AA^+b$.
$x = x_0 + x_n$ where $x_0 = \pi_{\mathcal{N}(A)^\perp}(x)$ and $x_N = \pi_{\mathcal{N}(A)}(x)$ Then we see did that we can vary x by changing its component on the null space x_n but we claim that there is a unique x_N
$A^+ A x = A^+ A A^+ b$
$A^+ A x = A^+ A x_0 + A^+ A x_N$ so the last part = 0
$A^+ A x_0 = \pi_{\mathcal{N}(A)^\perp}(x_0) = x_0$ therefore unique
$\implies$ If $Ax = AA^+ b$, then $x_0 = A^+ b$

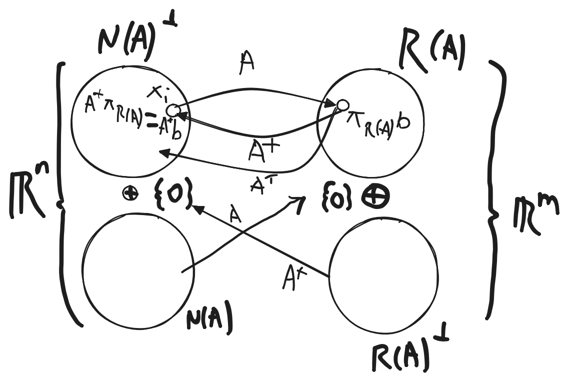
$\implies$ Then all solutions to

$$\min_x \|Ax - b\|_2^2$$

have form: $x^* = A^+b + x_N$ where $x_N \in \mathcal{N}(A)$

How to find solution of minimum norm:

Pythagorean Theorem: $\|A^+b + x_N\|_{2^2} = \|A^+b\|_2^2 + \|x_N\|_2^2$
$\|A^+b\|_2^2 + \|x_N\|_2^2 \geq \|A^+b\|_2^2$ w/equality $\iff$ $x_N = 0$
Means min norm solution is $x^* = A^+b$



Variations on Least Square:

Linearly Constrained LS

$$\min_x \|Ax - y\|^2 \text{ s.t. } Cx = d$$

Represents flow in = flow out in a network essentially a real life constraint.
If problem is feasible, then consider: $x_0 = C^+d$, is a solution to $Cx = d$ we can then say that all solutions to $Cx = d$ is of form $x = x_0 + Nz$ where columns of N form basis for the $\mathcal{N}(C)$.

This changes the problem to

$$\min_z \|A(x_0 + Nz) - y\|_2^2 = \min_z \|ANz - (y - AC^+d)\|_2^2$$

We can just rewrite AN to A' and $y - AC^+d$ to be some vector b and look at it like a LS problem therefore the min norm solution is
$z^* = (AN)^+(y - AC^+d)$
With all solutions including any additional vector in $\mathcal{N}(AN)$

Least Squares when Different Hilbert Norm

Off Script
Least squares when $\|x\|^2 = \langle x, x\rangle$ for some inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^m$
Recall: There exists positive definite W s.t. $\|y\|^2 = y^\top W y$ and that p.d. matrices can be split into $(W^{1/2})^\top(W^{1/2})$ where $W^{1/2}$ is also p.d.

$$\begin{aligned}
\min_x \|Ax-b\|^2 &= (Ax-b)^\top W(Ax-b) \\
&= (Ax-b)^\top (W^{1/2})^\top (W^{1/2})(Ax-b) \\
&= (W^{1/2}Ax - W^{1/2}b)^\top (W^{1/2}Ax - W^{1/2}b) \\
&= \|W^{1/2}Ax - W^{1/2}b\|_2^2 \\
&\implies \ x^* = (W^{1/2}A)^+ W^{1/2}b + \mathcal{N}(A)
\end{aligned}$$

He said this would be a good exam problem 💀. Since W is invertible and p.d. The null space Is the same as A. If W was p.s.d instead (which cannot be used to represent norm due to positive definiteness) then the null space would be $\mathcal{N}(W^{1/2}A)$.

Weighted Least Squares

$$\min_x \sum_i w_i |a_i^\top x - b_i|^2, \ w_i > 0$$

The above was a special case of previous where $W = \operatorname{diag}(w_1, \ldots, w_m)$. The weights can't be negative as then it's not a convex problem and the minimum maybe $-\infty$

ℓ^2 regularised LS

From Lecture 1

$$\min_x \|Ax-b\|_2^2 + \lambda^2\|x-x_0\|_2^2 \quad \lambda \geq 0$$

It's called regularised as the last term is a regularised. As λ increases in encourages values that are closer to x_0.
We can rewrite it as

$$\min_x \| \begin{bmatrix} A \\ \lambda I \end{bmatrix} x - \begin{bmatrix} b \\ \lambda x_0 \end{bmatrix} \|_2^2$$

Since ℓ^2 norm is separable

Question: What about e.g., ℓ^1 regularisation (**LASSO**)
Useful as it encourages sparse solutions (a lot of 0s) useful compressed sensing. As the cost

function means that it wants to be closer to 0 as let's say we put it at 0.1 then we get 0.1 for ℓ^1 whereas for ℓ^2 we would get 0.01 therefore encouraging 0 more for ℓ^1.
You **cannot** transform this into a least squares problem

Examples

"*Time Series Analysis*" of Autoregressive model.

$$y(k) = w_1 y(k-1) + w_2 y(k-2) \dots w_n y(k-n) + e(k)$$

where $e(k)$ is an error term
Suppose, we know $(y(k-1) \dots y(k-n))^\top$ and can write it as vector $\phi(k)$ and weights $(w_1 \dots w_n) = w^\top$ then best linear estimate is $w^\top \phi(k) \approx y(k) \implies$ estimation error $e(k) = y(k) - w^\top \phi(k)$.

Suppose we don't know model w, but have data $\phi(1), \dots, \phi(N)$How to estimate W?

For next lecture.