Solvability of least squares problems

Linear systems: $A\mathbf{x} = \mathbf{y}$ with A square: unique solution if A nonsingular

Linear least squares problems: $\min\|A\mathbf{x} - \mathbf{y}\|$ with A tall thin: unique solution if...?

Example:

$$\min\|A\mathbf{x} - \mathbf{y}\|, \quad A = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 1 & 3 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0 \\ 3 \\ 1 \\ 2 \end{bmatrix}.$$

Solution: We can 'match' the first three entries (but not the 4th).
$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ solves the problem. But also $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$. Or $\mathbf{x} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$...

Full column rank definition

What is going on: there is a vector $\mathbf{z} \neq 0$ in $\ker A$: $A \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = 0$.

If $\mathbf{x}$ is a solution, then so is $\mathbf{x} + \mathbf{z}, \mathbf{x} + 2\mathbf{z}, \mathbf{x} - 37\mathbf{z} \dots$

Definition

We say that $A \in \mathbb{R}^{m \times n}$ has full column rank if $\ker A = \{0\}$, or, equivalently: $\operatorname{rank} A = n$, or, equivalently: there is no $\mathbf{z} \in \mathbb{R}^n, \mathbf{z} \neq 0$ such that $A\mathbf{z} = 0$.

We shall see, via several equivalent conditions, that the least squares problem $\min \|A\mathbf{x} - \mathbf{y}\|$ has a unique solution if and only if A has full column rank.

Criterion for full column rank

Theorem

A has full column rank if and only if $A^T A$ is positive definite.

We already saw (lecture on orthogonal matrices) that $A^T A$ is symmetric and positive semidefinite.

For each $\mathbf{z} \neq 0$, $\mathbf{z}^T A^T A\mathbf{z} = \|A\mathbf{z}\|^2 \geq 0$.

Proof: A full column rank $\iff$ $A\mathbf{z} \neq 0$ for all $\mathbf{z} \neq 0$ $\iff$ $\mathbf{z}^T A^T A\mathbf{z} = \|A\mathbf{z}\|^2 \neq 0$ for all $\mathbf{z} \neq 0$

We can test the matrix from our earlier example, using eigenvalues.

```
>> A = [1 -1 0; 2 1 3; 1 0 1; 0 0 0];
>> eig(A'*A)
ans =
   2.6232e-16
   2.0718e+00
   1.5928e+01
```

Least squares problems — solution

Suppose A has full column rank. Then $\min\|A\mathbf{x}-\mathbf{y}\|$ can also be written as

$$\begin{aligned}
&\min_{\mathbf{x}\in\mathbb{R}^n}\frac{1}{2}\|A\mathbf{x}-\mathbf{y}\|^2=\min_{\mathbf{x}\in\mathbb{R}^n}\frac{1}{2}(A\mathbf{x}-\mathbf{y})^T(A\mathbf{x}-\mathbf{y})\\
&=\min_{\mathbf{x}\in\mathbb{R}^n}\frac{1}{2}\left(\mathbf{x}^TA^TA\mathbf{x}-\mathbf{y}^TA\mathbf{y}-\mathbf{x}^TA^T\mathbf{y}+\mathbf{y}^T\mathbf{y}\right)\\
&=\min_{\mathbf{x}\in\mathbb{R}^n}\frac{1}{2}\mathbf{x}^TA^TA\,\mathbf{x}-\mathbf{y}^TA\mathbf{x}+\frac{1}{2}\mathbf{y}^T\mathbf{y}
\end{aligned}$$

We have transformed the problem into the one of finding the minimum of a quadratic function $f(\mathbf{x})$ — sounds familiar?

Some optimization

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \mathbf{x}^T A^T A \mathbf{x} - \mathbf{y}^T A \mathbf{x} + \frac{1}{2} \mathbf{y}^T \mathbf{y}$$

Gradient $A^T A \mathbf{x} - A^T \mathbf{y}$.
Hessian $A^T A \succ 0$. $\rightarrow$ strictly convex!

The minimum exists unique, and can be found with

$$0 = \text{gradient} = A^T A \mathbf{x} - A^T \mathbf{y},$$

or

$$A^T A \mathbf{x} = A^T \mathbf{y}.$$

$A^T A$ is square invertible (because it's positive definite), so this linear system has a unique solution.

Can be solved with many methods: Gaussian elimination, LU factorization, QR (you'll see it soon),...

Computational cost

If done naively: (for $A \in \mathbb{R}^{m \times n}$, $m > n$, ignoring lower-order terms)

1. Computing $A^T A$: $2mn^2$.
2. Computing $A^T \mathbf{y}$: $2mn$ (lower-order).
3. Solving $A^T A \mathbf{x} = A^T \mathbf{y}$ with Gaussian elimination / LU factorization: $\frac{2}{3} n^3$.

Trick 1 using symmetry, we can skip half of the entries of $A^T A$.
Trick 2 a better way to solve linear systems with posdef matrices, Cholesky factorization, $A^T A = R^T R$ (we'll see it later).

1. Computing $A^T A$: mn^2.
2. Computing $A^T \mathbf{y}$: $2mn$ (lower-order).
3. Solving $A^T A \mathbf{x} = A^T \mathbf{y}$ with Cholesky: $\frac{1}{3} n^3$.

Geometric idea

TL;DR: can't solve $A\mathbf{x} = \mathbf{y}$? Multiply both sides by A^T and try again!

Geometric idea The residual $A\mathbf{x} - \mathbf{y}$ is orthogonal to any vector $A\mathbf{v} \in \operatorname{span} A$: $(A\mathbf{v})^T(A\mathbf{x} - \mathbf{y}) = 0$.

This method to solve LS problems is known as method of normal equations ('normal' is a fancy word for 'perpendicular/orthogonal').

Pseudoinverse

We showed that the solution of $\min\|A\mathbf{x} - \mathbf{y}\|$ is given by

$$\mathbf{x}_* = (A^T A)^{-1} A^T \mathbf{y}$$

(if A has full column rank).

Definition

The (Moore-Penrose) pseudoinverse of a matrix A with full column rank is $A^+ := (A^T A)^{-1} A^T$.

So we can write $\mathbf{x} = A^+ \mathbf{y}$ for the solution of a LS problem.

This generalizes the concept of inverse A^{-1} to a non-square A.

Non-obvious consequence: the solution is always obtained by multiplying $\mathbf{y}$ by a certain matrix. In particular, the solution of $\min\|A\mathbf{x} - (\mathbf{y}_1 + \mathbf{y}_2)\|$ is the sum of the two solutions of $\min\|A\mathbf{x}_1 - \mathbf{y}_1\|$ and $\min\|A\mathbf{x}_2 - \mathbf{y}_2\|$.

Note that $A^+ A = I_n$, but $AA^+ \neq I_m$ (there is no matrix such that $AA^+ = I_m$, for rank reasons.)

The other side

Sometimes in ML the same problem is formulated with multiplications on the other side: $\mathbf{w} \in \mathbb{R}^{1\times n}$ row vector of unknown weights, $X \in \mathbb{R}^{n\times m}$ matrix with each "feature" as a row, $\mathbf{y} \in \mathbb{R}^{1\times m}$ target (row) vector:

$$\min_{\mathbf{w}} \|\mathbf{w}X - \mathbf{y}\|_2.$$

This is the same problem, apart from notation. If $X \in \mathbb{R}^{n\times m}$ is short-fat ($n \leq m$) with linearly independent rows, then its pseudoinverse is defined as

$$X^+ = X^T(XX^T)^{-1}.$$

(Mnemonic: you must invert a matrix with the small dimension as its side.)

Exercises

1. Can a short-fat matrix $A \in \mathbb{R}^{m\times n}$, $n > m$, have full column rank, i.e., $\operatorname{rk} A = n$?
2. Write $\mathbf{x} = \mathbf{x}_* + \mathbf{z}$, where $\mathbf{x}_* = (A^TA)^{-1}A^T\mathbf{y}$ and $\mathbf{z}$ is an arbitrary vector, and show with algebraic manipulations that

$$f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{y}\|^2 = \frac{1}{2}\mathbf{x}_*^T A^T A\mathbf{x}_* + \frac{1}{2}\mathbf{z}^T A^T A\mathbf{z} + \frac{1}{2}\mathbf{y}^T\mathbf{y}.$$

   Use this formula to give another proof that $\mathbf{x}_*$ is the solution of the minimum problem.
3. Take a simple linear least squares problem, e.g. $\min \left\| \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 1 \end{bmatrix} \mathbf{x} - \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\|^2$. Try to solve it numerically with Matlab using gradient descent, which you saw in prof. Frangioni's lectures, and compare the iterates $\mathbf{x}_k$ with the exact solution $\mathbf{x}_*$. How many iterations do you need to get within, for instance, 10^{-5} of the exact solution?

Book references: Trefethen-Bau, Lecture 11; Demmel, Sections 3.1, 3.2; Eldén, Section 3.6.