Lecture 1 — Solutions via Moore–Penrose and Tikhonov

Problem statement (from the class exercise)

We study the (overdetermined) system

$$\begin{cases} 2x + y = 5, \\ x - y = 1, \\ x + y = 2, \end{cases} \qquad \text{i.e. } Ax \approx b,\ A = \begin{bmatrix} 2 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix},\ b = \begin{bmatrix} 5 \\ 1 \\ 2 \end{bmatrix}.$$

The system is *inconsistent*; we seek principled approximations.

1 Moore–Penrose least–squares solution

Normal–equations view

For full column rank ($\operatorname{rank}(A) = 2$), the least–squares (LS) solution is

$$x_{\mathrm{LS}} \;=\; (A^\top A)^{-1} A^\top b \;=\; A^+ b.$$

For our A, b:

$$A^\top A = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}, \quad A^\top b = \begin{bmatrix} 13 \\ 6 \end{bmatrix}, \quad x_{\mathrm{LS}} = \begin{bmatrix} \frac{27}{14} \\ \frac{5}{7} \end{bmatrix} \approx \begin{bmatrix} 1.928571 \\ 0.714286 \end{bmatrix}.$$

Residual $r = b - Ax_{\mathrm{LS}} = \left[\frac{1}{14}, -\frac{5}{7}, \frac{1}{2}\right]^\top$ and satisfies $A^\top r = 0$ (orthogonality to $\operatorname{col}(A)$).

Geometric meaning

x_{LS} is the unique vector whose image Ax_{LS} is the orthogonal projection of b onto the column space $\operatorname{col}(A)$. Equivalently, $r = b - Ax_{\mathrm{LS}} \perp \operatorname{col}(A)$.

SVD view (numerically stable)

Let $A = U\Sigma V^\top$ (thin SVD, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2)$ with $\sigma_1 \geq \sigma_2 > 0$). Then

$$x_{\mathrm{LS}} = A^+ b \;=\; V\Sigma^+ U^\top b \quad \text{with} \quad \Sigma^+ = \mathrm{diag}\Big(\sigma_1^{-1}, \sigma_2^{-1}\Big).$$

This expresses LS as filtered components along right singular vectors; small singular values (ill-conditioning) amplify noise.

2 Tikhonov (ridge) regularization

To control variance / ill-conditioning, solve

$$x_\lambda \;=\; \arg\min_x \;\|Ax - b\|_2^2 + \lambda\|x\|_2^2 \quad (\lambda > 0),$$

with closed form

$$x_\lambda \;=\; (A^\top A + \lambda I)^{-1} A^\top b.$$

For our A, b:

$$A^\top A + \lambda I = \begin{bmatrix} 6+\lambda & 2 \\ 2 & 3+\lambda \end{bmatrix}, \quad x_\lambda = \frac{1}{(6+\lambda)(3+\lambda) - 4} \begin{bmatrix} 3+\lambda & -2 \\ -2 & 6+\lambda \end{bmatrix} \begin{bmatrix} 13 \\ 6 \end{bmatrix}.$$

Limits. $x_\lambda \to x_{\mathrm{LS}}$ as $\lambda \downarrow 0$; and $x_\lambda \to 0$ as $\lambda \uparrow \infty$ (with standard ℓ_2 penalty).

SVD filter factors. With $A = U\Sigma V^\top$,

$$x_\lambda \;=\; \sum_{i=1}^{2} \frac{\sigma_i}{\sigma_i^2 + \lambda} \langle b, u_i \rangle \, v_i,$$

so each singular direction is damped by $\sigma_i/(\sigma_i^2 + \lambda)$; small σ_i are strongly suppressed. This is closely related to truncated SVD (TSVD), which *zeros out* directions with tiny σ_i.

3 Other geometric possibilities (beyond ℓ_2)

- **Weighted LS:** $\min_x \|W(Ax - b)\|_2^2$ to privilege certain equations (e.g. measurement reliabilities).
- **ℓ_1 (LAD) fit:** $\min_x \|Ax - b\|_1$ (robust to outliers; gives a different projection geometry via polytope distance).
- **ℓ_∞ (Chebyshev) fit:** $\min_x \|Ax - b\|_\infty$ (minimax worst–case residual; intersection of strips around the lines).

- **Constrained LS:** add box/convex constraints (e.g. nonnegativity) to encode prior knowledge.
- **Geometric "multi-line" compromise:** each equation is a line; LS chooses the point where the vector of signed distances is orthogonal to col(A); ℓ_1 chooses a point minimizing the sum of absolute distances; ℓ_∞ minimizes the maximal distance.

4 Numerical summary for the class system

$$x_{\mathrm{LS}} = \begin{bmatrix} \frac{27}{14} \\ \frac{5}{7} \end{bmatrix} \approx \begin{bmatrix} 1.928571 \\ 0.714286 \end{bmatrix}, \qquad r = b - Ax_{\mathrm{LS}} = \begin{bmatrix} \frac{1}{14} \\ -\frac{5}{7} \\ \frac{1}{2} \end{bmatrix}, \quad A^\top r = \mathbf{0}.$$

$$x_\lambda = (A^\top A + \lambda I)^{-1} A^\top b \ \text{ (closed form above).}$$

Example values:

$$\lambda = 0.1: \ x_{0.1} \approx \begin{bmatrix} 1.920 \\ 0.709 \end{bmatrix}, \qquad \lambda = 1: \ x_1 \approx \begin{bmatrix} 1.875 \\ 0.688 \end{bmatrix}.$$

As λ grows, the fit relaxes and the norm of x_λ shrinks.

5 Computation in R

Data

```
A <- matrix(c(2,1,
              1,-1,
              1,1), nrow=3, byrow=TRUE)
b <- c(5,1,2)
```

Moore–Penrose via normal equations

```
x_LS <- solve(t(A) %*% A, t(A) %*% b)
# or, with SVD (more stable):
sv  <- svd(A)
x_LS_svd <- sv$v %*% (t(sv$u) %*% b / sv$d)
```

Moore–Penrose via pseudoinverse

```
# install.packages("MASS")
library(MASS)
x_MP <- ginv(A) %*% b
```

Tikhonov (ridge) regularization

```
ridge <- function(lambda) {
  solve(t(A) %*% A + lambda * diag(ncol(A)), t(A) %*% b)
}
x_lam_01 <- ridge(0.1)
x_lam_1  <- ridge(1.0)
```

Diagnostics

```
r   <- b - A %*% x_LS
t(A) %*% r      # should be (numerically) c(0,0)
norm(r, type="2")
```

6 When to use which?

- Use **LS / pseudoinverse** when the model is well-conditioned and you want the best ℓ_2 fit.
- Use **Tikhonov** when columns of A are nearly dependent or data are noisy; tune λ to trade bias for variance.
- Consider ℓ_1 or **weighted** fits for outliers / heterogeneous noise; **constraints** to reflect domain knowledge.