It is customary to rename each column vector $a_i^\top$ as x_i (where $x_i \in \mathbb{R}^n$) and to rename the input data matrix A as X, so that the row vector $x_i^\top$ are the *rows* of the $m \times n$ matrix X

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix}.$$

Our optimization problem, called *ridge regression*, is

Program (RR1):

$$\text{minimize} \quad \|y - Xw\|^2 + K\|w\|^2,$$

which by introducing the new variable $\xi = y - Xw$ can be rewritten as

Program (RR2):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K w^\top w \\ &\text{subject to} \\ &\qquad\qquad y - Xw = \xi, \end{aligned}$$

where $K > 0$ is some constant determining the influence of the regularizing term $w^\top w$, and we minimize over ξ and w.

The objective function of the first version of our minimization problem can be expressed as

$$\begin{aligned} J(w) &= \|y - Xw\|^2 + K\|w\|^2 \\ &= (y - Xw)^\top (y - Xw) + K w^\top w \\ &= y^\top y - 2w^\top X^\top y + w^\top X^\top X w + K w^\top w \\ &= w^\top (X^\top X + K I_n) w - 2w^\top X^\top y + y^\top y. \end{aligned}$$

The matrix $X^\top X$ is symmetric positive semidefinite and $K > 0$, so the matrix $X^\top X + K I_n$ is positive definite. It follows that

$$J(w) = w^\top (X^\top X + K I_n) w - 2w^\top X^\top y + y^\top y$$

is strictly convex, so by Theorem 40.13(2)-(4), it has a unique minimum iff $\nabla J_w = 0$. Since

$$\nabla J_w = 2(X^\top X + K I_n) w - 2X^\top y,$$

we deduce that

$$w = (X^\top X + K I_n)^{-1} X^\top y. \tag{$*_{wp}$}$$

There is an interesting connection between the matrix $(X^\top X + K I_n)^{-1} X^\top$ and the pseudo-inverse X^+ of X.