$Ax = b$ make use of a factorization of A (QR decomposition, SVD decomposition), using orthogonal matrices defined next.

Given an $m \times n$ matrix $A = (a_{kl})$, the $n \times m$ matrix $A^\top = (a^\top_{ij})$ whose ith row is the ith column of A, which means that $a^\top_{ij} = a_{ji}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, is called the *transpose* of A. An $n \times n$ matrix Q such that

$$QQ^\top = Q^\top Q = I_n$$

is called an *orthogonal matrix*. Equivalently, the inverse Q^{-1} of an orthogonal matrix Q is equal to its transpose $Q^\top$. Orthogonal matrices play an important role. Geometrically, they correspond to linear transformation that preserve length. A major result of linear algebra states that every $m \times n$ matrix A can be written as

$$A = V \Sigma U^\top,$$

where V is an $m \times m$ orthogonal matrix, U is an $n \times n$ orthogonal matrix, and Σ is an $m \times n$ matrix whose only nonzero entries are nonnegative diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, where $p = \min(m, n)$, called the *singular values* of A. The factorization $A = V\Sigma U^\top$ is called a *singular decomposition* of A, or *SVD*.

The SVD can be used to "solve" a linear system $Ax = b$ where A is an $m \times n$ matrix, even when this system has no solution. This may happen when there are more equations that variables $(m > n)$, in which case the system is overdetermined.

Of course, there is no miracle, an unsolvable system has no solution. But we can look for a *good approximate solution*, namely a vector x that minimizes some measure of the error $Ax - b$. Legendre and Gauss used $\|Ax - b\|_2^2$, which is the squared Euclidean norm of the error. This quantity is differentiable, and it turns out that there is a unique vector x^+ of minimum Euclidean norm that minimizes $\|Ax - b\|_2^2$. Furthermore, x^+ is given by the expression $x^+ = A^+ b$, where A^+ is the *pseudo-inverse* of A, and A^+ can be computed from an SVD $A = V\Sigma U^\top$ of A. Indeed, $A^+ = U\Sigma^+ V^\top$, where Σ^+ is the matrix obtained from Σ by replacing every positive singular value σ_i by its inverse σ_i^{-1}, leaving all zero entries intact, and transposing.

Instead of searching for the vector of least Euclidean norm minimizing $\|Ax - b\|_2^2$, we can add the penalty term $K\|x\|_2^2$ (for some positive $K > 0$) to $\|Ax - b\|_2^2$ and minimize the quantity $\|Ax - b\|_2^2 + K\|x\|_2^2$. This approach is called *ridge regression*. It turns out that there is a unique minimizer x^+ given by $x^+ = (A^\top A + KI_n)^{-1} A^\top b$, as shown in the second volume.

Another approach is to replace the penalty term $K\|x\|_2^2$ by $K\|x\|_1$, where $\|x\|_1 = |x_1| + \cdots + |x_n|$ (the ℓ^1-norm of x). The remarkable fact is that the minimizers x of $\|Ax - b\|_2^2 + K\|x\|_1$ tend to be *sparse*, which means that many components of x are equal to zero. This approach known as *lasso* is popular in machine learning and will be discussed in the second volume.