In particular, the flow of a vector field whose divergence is everywhere negative contracts volume.

Suppose that \( g : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R} \) is smooth and, for notational convenience, let \( dx = dx_1 dx_2 \cdots dx_n \). The (Reynolds) transport theorem states that

\[
\frac{d}{dt} \int_{\phi_t(\Omega)} g(x, t) \, dx = \int_{\phi_t(\Omega)} g_t(x, t) + \text{div}(gf)(x, t) \, dx.
\]

The proof is elementary (see for example [18]).

## A.11 Least Squares and Singular Value Decomposition

The basic problem of linear algebra is to solve for the unknown vector \( x \) in the system of linear equations \( Ax = b \), where \( A \) is a matrix and \( b \) is a vector. In case \( A \) is a square matrix that is nonsingular (its determinant is not zero or its columns are linearly independent), there is a unique solution \( x = A^{-1}b \).

In general, the worst possible way to compute the solution \( x \) is to compute the matrix inverse. Two central principles of numerical linear algebra state: never compute a determinant and never compute the inverse of a matrix. The best solution methods for linear systems are based on Gaussian elimination or iteration. Some iterative methods are discussed in this book.

There are important problems in applied mathematics that require a solution of \( Ax = b \) in case \( A \) is not a square matrix or \( A \) is singular. The prime example is linear regression where the matrix is generally not square: We are given a finite set of points in the plane \( (x_i, y_i) \) for \( i = 1, 2, 3, \ldots, N \) and asked to find the best fitting line. More precisely, the problem is to find the line such that the sum of the squares of the deviations at the \( x_i \) from the line to \( y_i \) is minimized, that is,

\[
\min_{(m, b) \in \mathbb{R}^2} \sum_{i=1}^{N} |mx_i + b - y_i|^2.
\]

This problem may be recast into the abstract form

\[
\min_{x \in \mathbb{R}} |Ax - b|^2;
\]

where the vertical bars denote the Euclidean norm; \( A \) is the \( N \times 2 \)-matrix with first column the transpose of the row vector \( (x_1, x_2, x_3, \ldots, x_N) \) and
second column the $N$-vector all of whose entries are one; $b$ is the transpose of the row vector $(y_1, y_2, y_3, \ldots, y_N)$; and $x$ is the column vector $(m, b)$.

The corresponding matrix equation $Ax = b$ is a prototypical example of an overdetermined system of linear equations (more equations than unknowns). This equation has a solution exactly when every coordinate pair $(x_i, y_i)$ lies on the same line. The purpose of linear regression is to find a line that best fits the data when the data points do not all lie on the same line.

The reason for the squares is to simplify the mathematics. For instance, the problem $\sum_{i=1}^{N} \min_{m,b} |mx_i + b - y_i|$ is more difficult. The key point is that the Euclidean norm is defined by an inner product. In fact, using the usual inner (dot) product $\langle v, w \rangle := \sum v_i w_i$, the length of a vector $v$ is defined to be $|v| = \sqrt{\langle v, v \rangle}$. The square of the length is just the inner product.

Assume that $A$ is $m \times n$ with $m \geq n$ and define $f : \mathbb{R}^n \rightarrow \mathbb{R}$ by $f(x) = |Ax - b|^2 = \langle Ax - b, Ax - b \rangle$. We wish to minimize $f$ over all of $\mathbb{R}^n$. From calculus we know that the derivative of $f$ must vanish at a minimum. To compute the derivative $Df(x)$ abstractly, which is best in this case, recall that $Df(x)$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}$. Let $v$ denote an arbitrary vector in $\mathbb{R}^n$. By the chain rule,

$$\frac{d}{dt} f(x + tv) \bigg|_{t=0} = Df(x)v.$$  

Thus, we have that

$$Df(x)v = \frac{d}{dt} \langle A(x + tv) - b, A(x + tv) - b \rangle \bigg|_{t=0}$$

$$= \langle Av, Ax - b \rangle + \langle Ax - b, Av \rangle$$

$$= 2 \langle Ax - b, Av \rangle.$$  

Let us denote the matrix transpose (interchanging rows and columns) of $A$ by $A^T$. An important property (which is easy to check by a computation in components) is that $\langle w, Az \rangle = \langle A^T w, z \rangle$ for every pair of vectors $w$ and $z$. (Alternatively, we may define the matrix transpose of $A$ to be the unique matrix $A^T$ that satisfies the inner product identity and then prove that in coordinates $A^T$ is obtained from $A$ by interchanging its rows and columns). Using the transpose, it follows that

$$Df(x)v = 2 \langle A^T Ax - A^T b, v \rangle.$$
Suppose that $\langle u, v \rangle = 0$ for every $v$. Then, in particular, $|u|^2 = \langle u, u \rangle = 0$ and $u = 0$. Hence, if the derivative $Df(x)$ is the zero matrix, then

$$A^T Ax - A^T b = 0.$$  

This latter equation is called the normal equation for the least squares problem. Since $A$ is $m \times n$, the matrix $A^T A$ is $n \times n$.

For simplicity, let us make an additional assumption: The columns of $A$ are linearly independent. In this case, $A^T A$ is invertible. To prove this fact, suppose that $v$ is a vector and $A^T Av = 0$. Taking the inner product with respect to $v$, we have that

$$0 = \langle A^T Av, v \rangle = \langle Av, Av \rangle = |Av|^2.$$  

Since the columns of $A$ are linearly independent and $Av$ is a linear combination of these columns, $Av = 0$ only if $v = 0$. This means that $A^T A$ has linearly independent columns and hence is invertible.

Under our assumption that the columns of $A$ are independent, the normal equation has a unique solution

$$x = (A^T A)^{-1} A^T b.$$  

Thus, our function $f$ has a unique critical point, which must be a minimum because $f(x) \geq 0$ for every $x$ and $f(x)$ grows to infinity as $|x|$ grows without bound.

The quantity $(A^T A)^{-1} A^T$ is called the pseudo inverse of $A$. Using the pseudo inverse, the linear regression problem is solved: the transpose of the vector $(m, b)$ is exactly $(A^T A)^{-1} A^T b$ for the given $N \times 2$ matrix $A$ and the $N$-vector $b$.

In practice, the pseudo inverse is not computed directly. The normal equations are solved by elimination or an iterative method; or, better yet, the pseudo inverse is computed using the singular value decomposition of $A$. The reason for not simply solving the normal equations is that these equations may be ill conditioned, for example the matrix $A^T A$ may be nearly singular.

The singular value decomposition of an $m \times n$ real matrix $A$, which always exists, has the form $A = U \Sigma V^T$, where $U$ is an $m \times m$ orthogonal matrix (that is, $U^T U = UU^T = I$), $\Sigma$ is an $m \times n$ diagonal matrix, and $V$ is an $n \times n$ orthogonal matrix. The square roots of the diagonal elements of $\Sigma$ are called the singular values of $A$. 
Using the singular value decomposition and the properties of its factors, the normal equation may be written in the form

\[ V \Sigma^T U^T \Sigma V^T x = V \Sigma^T U^T b, \]

and simplified to

\[ \Sigma^T \Sigma V^T x = \Sigma^T U^T b. \]

In case the matrix \( A \) has linearly independent columns, its singular values are all positive. The matrix \( \Sigma^T \Sigma \) is square and its diagonal elements are the squares of the singular values of \( A \). Thus, the inverse of \( \Sigma^T \Sigma \) is diagonal with diagonal elements the reciprocals of its diagonal elements. An easy calculation shows that \( \Sigma^+ := (\Sigma^T \Sigma)^{-1} \) is diagonal with diagonal elements the reciprocals of the singular values of \( A \). It follows that the least squares minimum is achieved at

\[ x = V \Sigma^+ U^T b. \]

The matrix \( V \Sigma^+ U^T \) is the singular value decomposition pseudo inverse of \( A \). The ease of the inversion of \( \Sigma^T \Sigma \) and the efficiency of the numerical algorithms available to calculate the singular value decomposition make the method presented here the most used numerical method for computation of least squares problems. Of course, the singular value decomposition has many other applications.

Algorithms for the numerical computation of the singular value decomposition are presented in all books on numerical linear algebra.

### A.12 The Morse Lemma

If a class \( C^\infty \) function \( f : \mathbb{R}^n \to \mathbb{R} \) has a nondegenerate critical point \( a \) (that is, \( Df(a) = 0 \) and zero is not an eigenvalue of the symmetric linear transformation representing the quadratic form \( D^2 f(a)(x,x) \)), then there is a \( C^\infty \) change of coordinates defined in a neighborhood of \( a \) that transforms \( f \) to the function \( \xi \mapsto f(a) + D^2 f(a)(\xi, \xi) \).

A general proof of Morse’s lemma is given in [6]. The proof for the one-dimensional case is elementary. Reduce to the case where the function \( f \) is given by \( f(x) = x^2 h(x) \) and \( h(0) > 0 \). Define \( g(x) = x \sqrt{h(x)} \) and prove that there is a function \( k \) such that \( g(k(x)) = x \). The desired change of coordinates is given by \( x = k(z) \).