Let's return to the least squares problem that we discussed earlier in the course, and go a level deeper in our understanding. We'll again look at a few different versions of least squares problems. We'll start with the total least squares problem, which the one we'll use most in the rest of the course. This will lead us into the SVD (singular value decomposition). Then we will revisit the version 1 least squares problem, and have a deeper look at it.

Version 2: (total least squares)

An example we saw earlier was fitting a line to a set of points in the plane, where we were minimizing the perpendicular distance to the line.

Given an $m \times n$ matrix A, where $m \geq n$, find a unit length vector u that minimizes $\| \mathbf{Au} \|$.

Note that if we don't restrict the minimization to be for $\mathbf{u}$ of unit length, then the minimum is achieved when $\mathbf{u} = \mathbf{0}$ which would be trivial.

We are trying to find the unit vector $\mathbf{u}$ that minimizes $\mathbf{u}^T\mathbf{A}^T\mathbf{Au}$. The matrix $\mathbf{A}^T\mathbf{A}$ is $n \times n$, symmetric, and positive semi-definite,[1] *i.e.* $\mathbf{u}^T\mathbf{A}^T\mathbf{Au} \geq 0$ whenever $\mathbf{u}$ is non-zero. From these properties one can show using basic linear algebra[2] that $\mathbf{A}^T\mathbf{A}$ has an orthonormal and complete set of eigenvectors. We will denote by $\mathbf{v}_i$ where $i = 1, \ldots, n$.

Any vector $\mathbf{u}$ can be written as a sum of these eigenvectors, namely $\mathbf{u} = \sum_i a_i \mathbf{v}_i$. Then

$$\mathbf{u}^T\mathbf{A}^T\mathbf{Au} = \sum_i a_i \mathbf{v_i}^T\mathbf{A}^T\mathbf{Av_i} = \sum_i a_i \lambda_\mathbf{i}$$

This quantity $\| \mathbf{Au} \|_2^2$ is minimized when λ_k is the smallest eigenvalue and $a_k = 1$ and $a_i = 0$ for all i such that $i \neq k$. That is, the unit length vector $\mathbf{u}$ that minimizes $\mathbf{u}^T\mathbf{A}^T\mathbf{Au}$ is the eigenvector of $\mathbf{A}^T\mathbf{A}$ with smallest eigenvalue.

SVD (Singular Value Decomposition)

One can solve such a minimization problem by finding the eigenvectors and eigenvalues of $\mathbf{A}^T\mathbf{A}$. Another approach is as follows. This will give us more intuition about what the matrix $\mathbf{A}$ does.

Let $\mathbf{V}$ be an $n \times n$ matrix whose columns are orthonormal eigenvectors $\mathbf{v}_i$ of $\mathbf{A}^T\mathbf{A}$. We can order them in descending order of their eigenvalues. Since the eigenvalues λ_i are non-negative, we can write them as $\lambda_i = \sigma_i^2$, that is, $\sigma_i = \sqrt{\lambda_i}$. We can define the $n \times n$ diagonal matrix $\mathbf{\Sigma}$ such that $\Sigma_{ii} = \sigma_i$ on the diagonal. The elements σ_i are called the *singular values* of $\mathbf{A}$. Note that $\mathbf{\Sigma}^2$ is an $n \times n$ diagonal matrix, and since $\mathbf{V}$ is defined have the eigenvectors in its columns, we have

$$\mathbf{A}^T\mathbf{AV} = \mathbf{V}\mathbf{\Sigma}^2.$$

Multiplying on the left by $\mathbf{V}^T$ gives:

$$\mathbf{V}^T\mathbf{A}^T\mathbf{AV} = \mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^2 = \mathbf{\Sigma}^2.$$

[1] It is obvious that $\mathbf{A}^T\mathbf{A}$ is symmetric, and it is easy to see that the eigenvalues of $\mathbf{A}^T\mathbf{A}$ are non-negative since $\mathbf{u}^T\mathbf{A}^T\mathbf{Au} \geq 0$ for any real $\mathbf{u}$.

[2] See Theorem 3 of here

This implies that

$$(\mathbf{AV})^T\mathbf{AV} = \mathbf{\Sigma}^2.$$

which means that the columns of $\mathbf{AV}$ are orthogonal and the ith column of $\mathbf{AV}$ has length σ_i.

Next, define a matrix $m \times n$ matrix $\mathbf{U}$ whose columns are orthonormal, namely they are the normalized columns of $\mathbf{AV}$:

$$\mathbf{U\Sigma} = \mathbf{AV} \tag{1}$$

Right multiplying by $\mathbf{V}^T$ gives:

$$\mathbf{A} = \mathbf{U\Sigma V^T}.$$

This is called the *singular value decomposition* of $\mathbf{A}$.

There are methods for computing the SVD of $\mathbf{A}$. Once you have the matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{\Sigma}$ you can choose the eigenvector (column of $\mathbf{V}$) with smallest eigenvalue. This give us the solution to the version 2 least squares problem.

Notes: The singular value decomposition gives us a geometric intuition for the $\mathbf{A}$ does when it maps a vector $\mathbf{x}$ in $\Re_n$ to a vector $\mathbf{Ax}$ in $\Re_m$. First, it rotates $\mathbf{x}$ in $\Re_n$ by multiplying by $\mathbf{V}^T$, namely by taking the inner product with each of the columns of $\mathbf{V}$. Then, it scales each of its coefficients in this new coordinate system by multiplying by the singular values. Finally, it uses the coefficients as weights in a sum of the n columns of $\mathbf{U}$, where these column vectors are in $\Re^m$.

- One often defines the singular value decomposition of $\mathbf{A}$ slightly differently than this, namely one defines the $\mathbf{U}$ to be $m \times m$, by just adding $m - n$ orthonormal columns. One also needs to add $m - n$ rows of 0's to Σ to make it $m \times n$, giving

  $$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m}\mathbf{\Sigma}_{m\times n}\mathbf{V}^T_{n\times n}$$

  For our purposes there is no important difference between these two decompositions.

- Matlab has a function `svd` which computes the singular value decomposition.

  ```
  [U, S, V] = svd(A).
  ```

- What if the columns of A are not independent? In this case, $\mathbf{A^TA}$ will have at least one eigenvalue $\lambda_i = 0$, and hence at least one $\sigma_i = 0$.

Version 1: (linear regression)

Earlier in the course, we discussed another version of the least squares problem. This comes up in many *non-linear* least squares problems, where one designs an approximate solution by linearizing the problem. This is what we did with the Lucas-Kanade algorithm, for example.

Here I will restate the problem, and then discuss the solution more deeply than what I did in lecture 5. Finally, I will bring in the singular value decomposition into the discussion.

Given an $m \times n$ matrix $\mathbf{A}$ with $m \geq n$ and a non-zero m-vector, $\mathbf{b} \neq \mathbf{0}$, minimize $\| \mathbf{Au} - \mathbf{b} \|_2$.

Note that $\| \cdot \|_2$ is the L_2 norm. Minimizing the L_2 norm is equivalent to minimizing the sum of squares of the elements of the vector $\mathbf{Au} - \mathbf{b}$, i.e. the L_2 norm is just the square root of the sum of squares.

In lecture 5, we gave a general solution to this problem, namely we derived the *normal equations*:

$$\mathbf{A}^T\mathbf{Au} = \mathbf{A}^T\mathbf{b}. \tag{2}$$

If the columns of the $m \times n$ matrix $\mathbf{A}$ are linearly independent then $\mathbf{A}^T\mathbf{A}$ is invertible[3], and so the solution is

$$\mathbf{u} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}. \tag{3}$$

The $n \times m$ matrix in Eq. (3) that maps $\mathbf{b}$ to $\mathbf{u}$ above is called the *pseudoinverse*[4] of $\mathbf{A}$:

$$\mathbf{A}^+ \equiv (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T.$$

By inspection,

$$\mathbf{A}^+\mathbf{A} = \mathbf{I}$$

and

$$\mathbf{AA}^+ = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T.$$

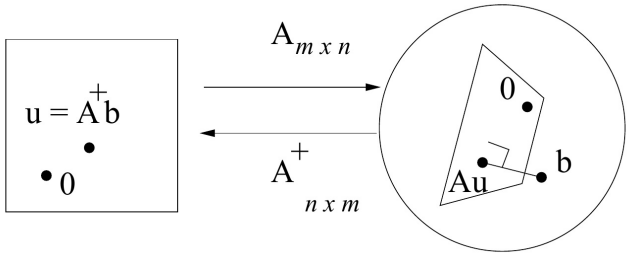
Note that the pseudoinverse maps in the reverse direction of $\mathbf{A}$, namely it maps $\mathbf{b}$ in a (higher dimensional) m-D space to some $\mathbf{u}$ in a (lower dimensional) n-D space.

What is the geometric interpretation of the pseudoinverse, in particular Eq. (2)? Since $\mathbf{b}$ is an m-dimensional vector and $\mathbf{A}$ is an $m \times n$ matrix, we can *uniquely* write $\mathbf{b}$ as a sum of a vector in the column space of $\mathbf{A}$ (that is, a sum of the column vectors of $\mathbf{A}$) and a vector in the space orthogonal to the column space of $\mathbf{A}$. To minimize $\| \mathbf{Au} - \mathbf{b} \|_2$, by definition we find the $\mathbf{u}$ such that the distance from $\mathbf{Au}$ to $\mathbf{b}$ is as small as possible. This is done by choosing $\mathbf{u}$ such that $\mathbf{Au}$ is the component of $\mathbf{b}$ that lies in the column space of $\mathbf{A}$, that is, $\mathbf{Au}$ is the orthogonal projection of $\mathbf{b}$ to the column space of $\mathbf{A}$. Note that if $\mathbf{b}$ already belonged in the column space of $\mathbf{A}$ then $\| \mathbf{Au} - \mathbf{b} \|$ would be 0 and there would be an exact solution.

Rather than inverting the mapping $\mathbf{A}$, the pseudoinverse $\mathbf{A}^+$ only inverts the component of $\mathbf{b}$ that belongs to the column space of $\mathbf{A}$, and it nulls out any component of $\mathbf{b}$ that is orthogonal to the column space of $\mathbf{A}$. Specifically, $\mathbf{AA}^+$ projects any vector $\mathbf{b} \in \Re^m$ onto the column space of $\mathbf{A}$ by removing from $\mathbf{b}$ the component that is orthogonal to the column space of $\mathbf{A}$. Note that the plane in the figure on the right contains the origin, since it includes $\mathbf{A}\ \mathbf{u} = \mathbf{0}$, where $\mathbf{u} = \mathbf{0}$. This property holds even when $\mathbf{A}$ doesn't have linearly independent columns.

[3] If $\mathbf{A}^T\mathbf{A}$ were not invertible, then it would mean that it there is non-zero vector $\mathbf{u}$ such that $\mathbf{A}^T\mathbf{Au} = \mathbf{0}$. This would mean that there is some linear combination of the columns of $\mathbf{A}$ (vector $\mathbf{A}\ \mathbf{u}$) that is orthogonal to each of the columns of $\mathbf{A}$, which is not possible since the columns of $\mathbf{A}$ are linearly independent.

[4] A more general definition of pseudoinverse is used in the case that the columns of $\mathbf{A}$ are not linearly independent. (Details omitted.)



Let's try to use the singular value decomposition to understand a bit more about what's happening here. If we substitute $\mathbf{U}\Sigma\mathbf{V}^T$ into the expression for $\mathbf{A}^+$, you we get

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathbf{T}}$$

where Σ^{-1} is a diagonal matrix with diagonal elements σ^{-1}. (See Exercises.) By similar substitution,

$$\mathbf{A}\mathbf{A}^+ = \mathbf{U}\mathbf{U}^{\mathbf{T}}$$

Notice that $\mathbf{U}\mathbf{U}^{\mathbf{T}}$ is not an identity matrix! The reason is that $\mathbf{U}$ is an $m \times n$ matrix with $m \geq n$, so $\mathbf{U}\mathbf{U}^{\mathbf{T}}$ in an $m \times m$ matrix. It has 1's on its first n diagonals. But it has 0's on its last $m - n$ diagonals. So multiplying $\mathbf{b}$ by $\mathbf{U}\mathbf{U}^{\mathbf{T}}$ is projecting $\mathbf{b}$ onto the column space (range) of $\mathbf{A}$, namely the space in $\Re^m$ spanned by the columns of $\mathbf{U}$.

