Some selected topics in linear algebra for SSI (3)
An Introduction to Inverse Problems

Prof. Brooks
Northeastern University
Spring 2011

This talk will relate to parts of
• Chapter 5
• Appendix B
Outline of Lecture on Inverse Problems

- What characterizes an inverse problem?
- A small-scale mathematical example
- The Singular Value Decomposition (SVD)
- Condition Number of a Matrix
- (Min-Norm) Least Squares via the SVD
- SVD Analysis of ill-posed problems
- Regularization of ill-posed problems
Inverse Problems

- What is an “Inverse Problem”?  
  - We measure the “output” of some observation system.  
  - We have a model of the input and the system, but one or both is  
    1. unknown, or  
    2. of a known form but with unknown parameters  
  - We’re interested in extracting  
    1. the unknown information, or  
    2. partial information about what’s unknown  
    from the measurements, using the model.

- Typically this is difficult because we lose information in the observation process
Some Example Inverse Problems

1. Low-pass filtered signal
2. Blurred or sensor-degraded image
3. Discretized integral or differential equation
   - Diffuse Optical Tomography
   - Electrical Impedance Tomography
   - Limited-View Straight-Ray Tomography
   - Geophysical problems: demining, buried waste localization, oil prospecting, . . .
   - Underwater coral reef characterization
   - Acoustic tomography
   - Cardiac Electrical Imaging
   - Functional Brain Mapping
   - . . .
Low-Pass Filter Example

How to restore shape of original signal?
Low-Pass Filtered Signal

- Observe output signal through known low-pass filter, want to reconstruct input
- Equivalently for linear system, input known and want to reconstruct filter
- If filter has no zeros:
  - Clearly inverse filter is high-pass; need to amplify high-frequency components in output
  - How to avoid amplifying noise too? Need to constrain solution
- What if filter has zeros? Need to modify inverse filter to compensate or need to constrain solution
Image Reconstruction

- Image is blurred due to movement or sensor imperfections
- Examples: Hubble telescope, Diffraction limit of lens (apodization), ...
- Distinct features in original image merged
- Have mathematical model for blur: may or may not be space-invariant (convolutional)
- Need to amplify small features in observation to recover original
- How to avoid amplifying noise too? Need to constrain solution
Problems Modeled by Integral or Differential Equations

• Observation is related to quantity or object of interest through integration with a smooth kernel representing a system.

• Examples: most useful subsurface problems: medical imaging, environmental imaging (see earlier list . . .)

• Attentuation, superposition, limited view problems all present
  – Attentuation: lose information in transit through subsurface medium (loss of SNR)
  – Superposition: distinct aspects of quantity or object are seen blurred together
  – Limited view: can’t see the object or quantity from all “angles” or “views”
  – Usually not space-invariant
  – Often finer discretization does not help, problem is fundamental: similarities (near-dependence) between observations
Examples

● Diffuse Optical Tomography:
  1. Loss of signal due to scattering and absorption from background, uninteresting inhomogeneities
  2. Blurring from scattering and multiple absorbers, uninteresting inhomogeneities
  3. Limited view because you can’t put sources and detectors where you’d like to (restricted to accessible skin surfaces)

● Limited-view Straight-Ray Tomography
  1. Loss of signal from scattering
  2. Blurring from scattering, imprecision, shot noise
  3. Limited view problem leads to matrix formulations
What Do We Want to Image?

- **Medium**: Passive Sensing Problems
  - Source comes from “natural” activity (we know or can characterize the source)
  - We observe it through an unknown medium and want to reconstruct or characterize the medium
  - Example: remote sensing using solar illumination

- **Medium**: Active Sensing Problems
  - We introduce energy into the medium
  - The energy interacts with the medium and we observe energy that emerges from the medium
  - Examples of “energy” used: optical, acoustics, Xray, magnetic, radiation, ...
  - We want to characterize
    1. medium’s physical structure (anatomy, earth layers)
    2. activity in medium (functional brain imaging, imaging with radioactive tracers)
What Do We Want to Image? (cont’d)

• **Source:** Intrinsic Source Problems
  – Source comes from “natural” activity
  – We observe it through a (known) medium and want to reconstruct or characterize the source
  – Examples: cardiac electrical imaging, brain electrical activity, earthquake characterization, astronomy

• **General Summary:** What we want to observe
  – Is weakened by the time (at the place) we observe it,
  – Has distinct features which are blurred together,
  – Or maybe even has features we don’t see at all (if system is singular and part of input is in its null space)
Linear Inverse Problems

• Assume problem is linear or linearizable:
  – Linear examples: straight-ray tomography, cardiac electrical imaging
  – Non-linear but linearizeable, e.g. via Born approximation

• So we have a linear model

\[ b = Ax + n \]

– \( b \) are our measurements,
– \( x \) is what we want to recover,
– \( A \) is our model of how these two are related
– \( n \) is measurement noise and perhaps model error

• Other non-linear approaches a research topic . . .
“Ill-conditioned” Forward Problems

• In most cases, the matrix governing the relationship between
  1. what we observe and
  2. what we are interested in

• has the following characteristics:
  1. It causes outputs to be more similar than inputs
  2. It has columns and rows which are close to being linearly dependent
  3. It requires amplification of some small components of output to
     recover many inputs

• We say it is “ill-conditioned” (will make this precise soon).

• Terminology:
  – **Forward Problem**: model of how to get from quantity of interest
    to measurements. Linear problem is *ill-conditioned*.
  – **Inverse Problem**: how to get from measurements and model to
    quantity of interest. We say it’s *ill-posed*.
  – Note: the two are interconnected.
Simple Numerical Example

Borrowed from P.C. Hansen:

• Let

\[
A = \begin{bmatrix}
0.16 & 0.10 \\
0.17 & 0.11 \\
2.02 & 1.29
\end{bmatrix}, \quad \text{and} \quad b_t = \begin{bmatrix} 0.27 \\ 0.28 \\ 3.31 \end{bmatrix}
\]

where \( b_t \) are “true” (noise-free) measurements.

• For this system the true solution \( x_t = [1.00 \ 1.00]^T \).

• Now suppose we observe

\[
b = \begin{bmatrix} 0.27 \\ 0.25 \\ 3.33 \end{bmatrix} = b_t + \epsilon,
\]

with

\[
\epsilon = \begin{bmatrix} 0.01 \\ -0.03 \\ 0.02 \end{bmatrix}.
\]

• Note that the length of \( b \) and of \( b_t \) differ by about 1%.
Least-Squares Solution

• The least-squares (LS) solution to this system is

\[ x_{LS} = \begin{bmatrix} 7.01 \\ -8.40 \end{bmatrix} \]

• Useless !! Almost no relation to the true value \([1.00 \ 1.00]^T \) !!

• Why? Well, it turns out that
  – the ratios of the elements of the two columns are all \( \approx 1.6 \) (1.60, 1.55, and 1.57).
  – Both true solution and LS solution “look” like true measurements:
    \( b_{LS} = Ax_{LS} = [0.282 \ 0.268 \ 3.324]^T \), compared to
    \( b_t = [0.26 \ 0.28 \ 3.31]^T \).
  – Two more bad solutions that look like the data:
    1. If we use \( x_1 = [1.65 \ 0]^T \) then
       \( b_1 = A x_1 = [0.264 \ 0.281 \ 3.333]^T \).
    2. If we set \( x_2 = [0 \ 2.58]^T \) then
       \( b_2 = A x_2 = [0.258 \ 0.284 \ 3.328]^T \).
Summarizing the situation

- In each case
  1. the residuals (difference between actual and predicted measurements, $b, b_{LS}, b_1, b_2$) are relatively small, and similar to noise-free $b_t$, and yet
  2. the solutions are not close to the true solution
- It seems the only hope is if we
  1. know something else ahead of time (a priori) about the solution, and
  2. force our solution to pay attention to this a priori knowledge.
Imposing Some a priori Knowledge

- Note: $x_{LS}$ has norm $\gg$ norm of $x_t$ (almost a factor of 8 more).
- We might know ahead of time that this is not realistic. So try:

$$\min_x \| Ax - b \|_2, \text{subject to } \| x \|_2 < \alpha$$

- For various values of $\alpha$ (note true value is $\sqrt{2}$) we get

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$x^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>[0.08 0.05]</td>
</tr>
<tr>
<td>1.385</td>
<td>[1.17 0.74]</td>
</tr>
<tr>
<td>10</td>
<td>[6.51 -7.60]</td>
</tr>
</tbody>
</table>

- Conclusions:
  - LS solution useless
  - Small residual does not mean good solution
  - Additional constraints can help but need to be careful
Let’s analyze the problem more precisely

- The physical system makes distinct features smaller and less distinct.
- A linear model of this results in a matrix which makes measurements “similar” to each other.
- More precisely, the columns (for overdetermined systems) or rows (for underdetermined systems) seem to be almost linearly dependent.
- So we need an analysis tool that shows us this behavior of a matrix.
- This tool the **Singular Value Decomposition**.
Remember: The Best Thing in Linear Algebra, the SVD

It turns out that the single most useful tool in linear algebra for solving systems of equations is a rather non-obvious decomposition of a matrix:

- For any $m \times n$ matrix $A$, it turns out that we can write it as the product of three matrices:
  1. An orthogonal matrix $U$
  2. A diagonal matrix $\Sigma$, and
  3. An orthogonal matrix $V^T$:

\[
A = U\Sigma V^T.
\]

- If $m \neq n$ we have several options about how to play with the dimensions. We’ll use this one: $U$ is $m \times m$, $\Sigma$ is $m \times n$, and $V$ is $n \times n$.

- This is called the Singular Value Decomposition (SVD), and the (diagonal) entries of $\Sigma$ are called the “singular values” of $A$. 

Remember a definition and a property

- **Definition:** Orthogonal matrices: for a square matrix $Q$, if
  \[ Q^T Q = I \]
  then we say $Q$ is orthogonal.

- Note this means that $q_i^T q_j = \delta(i - j)$, where
  1. by $q_k$ we mean the $k$'th column of $Q$, and
  2. $\delta(i - j) = 1$ when $i = j$ and $0$ otherwise.

- **Property:** the 2-norm of a vector is not changed by multiplying by an orthogonal matrix:
  \[ \|Qx\|_2^2 = x^T Q^T Q x = \|x\|_2^2. \]

- Consider the ellipsoid interpretation of matrix multiplication: since the length does not change, multiplying by an orthogonal matrix corresponds to a pure rotation.
Remember: Geometric Interpretation of the SVD

- Remember that multiplying by orthogonal matrices corresponds to a rotation.
- So we can interpret the SVD as a rotation, a scaling, and another rotation:
A Closer Look at $\Sigma$

• Note that definition of the sizes of $U$, $\Sigma$, and $V$ means that

1. If $A$ is square ($m = n$), $\Sigma$ is square:

   \[
   \Sigma = \begin{bmatrix}
   \sigma_1 & 0 & \ldots & 0 & 0 \\
   0 & \sigma_2 & 0 & \ldots & 0 \\
   \vdots & & \ddots & & \vdots \\
   0 & \ldots & \sigma_i & \ldots & 0 \\
   \vdots & & & & \\
   0^T & & & \sigma_m \\
   \end{bmatrix}
   \]

2. If $A$ is overdetermined ($m > n$), $\Sigma$ has the form

   \[
   \begin{bmatrix}
   \Sigma_0 \\
   0, \\
   \end{bmatrix}
   \]

   with $\Sigma_0$ $n \times n$ and $0$ an $(m - n) \times n$ block of zeros.

3. If $A$ is underdetermined ($m < n$), $\Sigma$ has the form

   $[\Sigma_0 \ 0]$, with $\Sigma_0$ $m \times m$ and $0$ an $m \times (n - m)$ block of zeros.
Some Relevant Properties of the SVD

- The SVD has many many properties, uses, and interpretations: we’ll concentrate on the ones most relevant here:
  - How to use it to approximately solve singular systems.
  - How to use it to decide if a system is close to singular.
  - How to use it to analyze and get useful results when systems are close to losing full rank (column, row, or both).

- So let’s list a few properties:
  - The entries of $\Sigma$ are called the “singular values” of $A$.
  - They are always ordered as $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)}$.
  - The number of non-zero singular values is equal to the rank of $A$.
  - The rest of the singular values are equal to 0.
  - So if rank is $r$, we have
    \[ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots \sigma_{\min(m,n)} = 0. \]
Two More Useful Facts

• FACT 1: Multiplying a matrix by a diagonal matrix:
  1. **Post-multiplying** multiplies each column by the corresponding diagonal element (consider column-oriented multiplication)
  2. **Pre-multiplying** multiplies each row by the corresponding diagonal element (consider dot-product or transpose of post-multiplying).

• FACT 2: A useful identity: if \( A \) and \( B \) are any matrices and \( D \) is a diagonal matrix (assume they’re all \( n \times n \) for convenience), then

\[
ADB = \sum_{i=1}^{n} d_i a_i b_i^T,
\]

where \( a_i \) (\( b_i \)) are the columns of \( A \) (\( B \)), \( d_i \) the \( i \)'th diagonal element.

– Note that each term in the sum is an \( n \times n \) matrix formed by the outer product of a column and a row.
– Write it down for some simple examples to convince yourself.
– Each term in the sum is a rank one matrix (each row of the product is a different number times the row vector \( b_i^T \)).
– Applies directly to the SVD
Quantifying Close-to-Losing-Rank

• Up to now we have treated linear independence (or full rank) as a binary property . . . you got it or you don’t.

• What about the vectors $[1, 1, 1]^T$ and $[1, 1, 1.00000001]^T$. In principle they’re independent, but in practice they point awfully close to the same direction.

• The first thing we might think of to measure ’how close to singular’ is the determinant. However
  – The determinant gives the volume of a parallelepiped described by the matrix.
  – Thus if the matrix is very extended in one direction, even if it is very narrow in another, it can have a reasonable determinant but be close to losing rank.

• Much better to use quantity based on the singular values: the “condition number”
The Condition Number

• The condition number $\kappa$ is defined as the $\sigma_1/\sigma_r$: ratio of the largest to the smallest non-zero singular value.

• Geometric interpretation: gives a measure of eccentricity of ellipse.

• Has many other uses (eg quantify sensitivity to numerical error).

• Two kinds of high-condition-number matrices:
  1. Ones where there is a big jump between 'large' and 'small' SV's,
  2. Ones where the decay of SV's is gradual.

• In the former case, we can truncate the smaller ones if we want—gives "low-rank" approximation.

• In the latter case, life is more difficult . . . and more interesting.
A Numerical Example

Consider the matrix

\[
A = \begin{bmatrix}
10^6 & 0 \\
0 & 10^{-6}
\end{bmatrix}
\]

- It’s determinant is 1.
- It’s condition number is \(10^{12}\).
- Suppose we want to solve \(Ax = b\), for this \(A\) and \(b = [1, 1]^T\). So \(x = [10^{-6}, 10^6]^T\).
- Note that small errors in the 2nd entry of \(b\) will be greatly magnified in any inversion—more soon.
- Let’s do some algebra to see how to use the SVD for inverse problems. But first:
Evaluating the 'Goodness' of a System of Equations

• In some sense, $\kappa$ defines 'how good', or at least 'how bad', a system of equations is.
  – In the sense of defining 'how independent' or 'dependent' the equations are, or equivalently
  – 'how much new information' each measurement does or doesn't bring.

• No well-defined algorithm to choose a 'good set'.

• But $\kappa$ is a good quantification tool for deciding if a particular set is good, or comparing two sets.

• Becomes a tool, for instance, to choose whether a particular set of wavelengths for reflectance spectroscopy is a good set or not

• $\kappa$ close to 1 is perfect (an orthogonal matrix has condition number 1)

• $\kappa$ “large”? $10^3$? $10^6$? $10^9$? Depends partly on the size of the system.
Solving a non-Singular System with the SVD

• First let’s assume \( A \) is square \((m \times m)\) and full rank (not singular). Then, if \( Ax = b \)

\[
\begin{align*}
Ax &= b \\
U\Sigma V^T x &= b
\end{align*}
\]

• Note in this case that all diagonal entries of \( \Sigma \) are non-zero and thus its inverse exists and is diagonal with \((i, i)\) entry \(1/\sigma_i\).

• Now we premultiply first by \( U^T \), then by \( \Sigma^{-1} \), then by \( V \), using the orthogonality of \( U \) and \( V \), and finally FACT 2:

\[
\begin{align*}
\Sigma V^T x &= U^T b \\
V^T x &= \Sigma^{-1} U^T b \\
x &= V \Sigma^{-1} U^T b \\
x &= \sum_{i=1}^{m} \frac{1}{\sigma_i} v_i u_i^T b.
\end{align*}
\]
What Does This Do For Us?

• So this gives us another way of solving a linear system. It’s not obvious this is very useful.

• But now suppose that $A$ is singular, with rank $r < m$ (so that we have only $r$ non-zero singular values).

• It turns out that a least-squares (min-norm) solution for this rank-deficient case can be achieved if we simply truncate the previous sum at $r$:

$$x^* = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T.$$

• We can extend this idea to all cases of interest.
Matrix Version of Same Result

• First, we define, for a $n \times n$ diagonal matrix $\Sigma$, another $n \times n$ diagonal matrix $\Sigma^\dagger$:

$$
\Sigma = \begin{bmatrix}
\sigma_1 & 0 & \ldots & 0 & 0 \\
0 & \sigma_2 & 0 & \ldots & 0 \\
\sigma_1 & 0 & \ldots & \sigma_r & 0 \\
0 & \ldots & 0 & 0 & 0 \\
\end{bmatrix}, \quad \Sigma^\dagger = \begin{bmatrix}
1/\sigma_1 & 0 & \ldots & 0 & 0 \\
0 & 1/\sigma_2 & 0 & \ldots & 0 \\
0 & \ldots & 1/\sigma_r & 0 & 0 \\
0 & \ldots & 0 & 0 & 0 \\
\end{bmatrix}
$$

• Using FACT 2 in reverse we can write

$$
x^* = V \Sigma^\dagger U^T b.
$$

• The matrix multiplying $b$ here is called the “pseudo-inverse” of $A$ and is usually denoted $A^\dagger$.

• Note that when $A$ is invertible, $A^\dagger = A^{-1}$. 
Extending this Approach

We’ve considered a number of situations: assuming we have \( m \) measurements and \( n \) unknowns

- Square non-singular systems: rank \( = m = n \)
- Square singular systems: rank \( = r < m = n \)
- Over-determined, full column rank systems: rank \( = n < m \)
- Over-determined, column rank deficient systems: rank \( = r < n < m \)
- Under-determined, full row rank systems: rank \( = m < n \)
- Under-determined, row rank deficient systems: rank \( r < m < n \)

In all these cases, the SVD approach leads directly to the (min-norm) least-squares solution:

- truncate the sum at \( r \), or equivalently
- define an appropriate \( \Sigma^{†} \)
Let’s Consider the Full-Rank Overdetermined Case

Can we verify that

\[ x^* = (A^T A)^{-1} A^T b \]

is the same as

\[ V \Sigma^\dagger U^T b? \]

- Starting from \( A = U \Sigma V^T \), with \( \Sigma = [\Sigma_0 0]^T \) and \( \Sigma_0 \) full rank and invertible (since we assumed \( A \) is full column rank).

- Plugging in we get

\[
\begin{align*}
x^* &= (A^T A)^{-1} A^T b \\
&= \left( (U \Sigma V^T)^T U \Sigma V^T \right)^{-1} (U \Sigma V^T)^T b \\
&= \left( V \Sigma^T U^T U \Sigma V^T \right)^{-1} (V \Sigma^T U^T) b \\
&= \left( V \Sigma^T \Sigma V^T \right)^{-1} (V \Sigma^T U^T) b
\end{align*}
\]

where the third line above follows from the order-reversing property of the transpose of a product and the last line from the orthogonality of \( U \).
Continuing on . . .

• Now let’s examine the middle term inside the product of matrices that needs to be inverted:

\[
\Sigma^T \Sigma = [\Sigma_0 \ 0] \begin{bmatrix} \Sigma_0 \\ 0, \end{bmatrix} = \Sigma_0^2,
\]

where each diagonal entry of \( \Sigma_0^2 \) is the square of the corresponding singular value.

• So, using this result, the order-reversing property of the inverse of a product, and the orthogonality of \( V \), we get

\[
x^* = (V \Sigma_0^{-2} V^T V \Sigma_0^T U^T) b
= (V \Sigma_0^{-2} \Sigma_0^T U^T) b
= V \begin{bmatrix} \Sigma_0^{-1} \\ 0 \end{bmatrix} U^T b
= \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T b.
\]

• Similar analysis for the other cases on our list . . .
Computing the SVD

- **Always** use someone else’s program . . .

- Seriously, the SVD is not at all straightforward to compute — its utility seems proportional to its non-obviousness.

- Matlab commands
  - \([U, S, V] = \text{svd}(A)\); computes the SVD as we’ve described it.
  - \(S = \text{svd}(A)\); computes just the singular values.
  - \([U, S, V] = \text{svd}(A,0)\); computes the “economy” SVD — it just computes and returns the columns of \(U\) and \(V\) that correspond to non-zero singular values.
Some Other Uses of the SVD

Just listing a few here:

- Quantify the gain of a system,
- Analyze approximate inversions for systems with large condition and gradually-decaying singular values,
- Filter signal from noise via subspace approach,
- Build bases for column, row, null, and left null spaces,
- Approximate matrices by other “nearby” matrices
- A note: combines results from eigenvalue decomposition of $A^T A$ and $AA^T$. 
So Where Does This Leave Us?

• For any kind of system, we can find a least-squares (min-norm) solution, using the SVD:
  1. Square singular systems
  2. Overdetermined full (column) rank systems
  3. Overdetermined (column) rank deficient systems
  4. Underdetermined full (row) rank systems
  5. Underdetermined (row) rank deficient systems

• Unified framework, simply need to take care of zero diagonal entries of $\Sigma$ via truncation.

• What about ill-posed inverse problems, though?
SVD Analysis of Ill-Posed Inverse Problems

Let’s analyze the situation using the SVD:

- Overdetermined case: Start from
  1. \( b = Ax \)
  2. \( A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T \)
  3. \( A \) is \( m \times n \) with \( m \geq n \), and
  4. \( A \) is full column rank, but just barely . . .

- So then:
  \[
  \hat{x} = V\Sigma^{-1}U^T b = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T b
  \]

- Now suppose \( b = b_t + n \), with \( n \) some noise in our measurements. Then
  \[
  \hat{x} = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T b_t + \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T n
  \]
  \[
  = x_t + x_e
  \]

- Note that the \( \sigma_i \) get smaller as \( i \) gets bigger, so the \( 1/\sigma_i \) get larger.
### Algebraic Interpretation

- Usually $u_i$ (and $v_i$) have more sign changes (higher frequency) as $i \to n$
- So the high frequency components in $b$ (usually due disproportionally to $n$, since quantities of interest are usually relatively low-frequency) are amplified by small $\sigma_i$
- If $u_i^T n$ does not $\to 0$ as $i \to n$, (or $\to 0$ slower than $\sigma_i \to 0$),
- then $\Rightarrow x_e \gg x_t$, in other words the Least-Square solution is useless.
- Remember the least-squares solution from our $3 \times 2$ example
- Why does noise dominate? Consider, e.g. white noise. We expect it to have equal strength at all frequencies (that’s what white means), so it stays the same strength as $\sigma_i \to 0$. 
Geometric Interpretation

• Recall that, geometrically $A = U\Sigma V^T$ means
  – Rotation, scaling, rotation, and small $\sigma_i$ mean close to losing dimension (rank)
  – Going back means amplifying axes by reciprocal of $\sigma_i$
  – For large $i$
    * small $\sigma_i$, large $1/\sigma_i$
    * Correspond to high-frequency singular vectors, matches high-frequency components of input
    * So we amplify high-frequency inputs (noise, model error, ...).
Very different solutions “look” very similar when we “measure” them.

⇒ Very similar measurements result in wildly different inverse solutions.
Regularization: One Way to Approach Ill-Posed Problems

- Characteristics of Forward Model of Ill-Posed Inverse Problems:
  1. \( \sigma_{\text{max}} / \sigma_{\text{min}} \) large
  2. \( \sigma_i \)'s decay smoothly

- Basic Idea: modify problem to make it well-posed

- Two ways to conceptualize:
  - Idea 1: Replace ill-posed problem with "similar", safer (better-posed) problem.
  - Idea 2: Force solution to be reasonable by restricting possible solutions (eg by size or smoothness), find closest solution in this restricted set of solutions

- The turn out to be closely related
Simple Idea: Truncated SVD (TVSD)

- **Idea 1**: Replace $A$ by closest (in 2 norm sense) well-conditioned approximation

- Do this by dropping small $\sigma_i$'s before inversion

$$\hat{x}_{\text{TSVD}} = \frac{1}{\sigma_i} v_i u_i^T b$$

for $p < n$

- In effect, restricts size $\| \hat{x} \|_2$ since

$$\hat{x}_{\text{TSVD}} = V^T \hat{\Sigma}^\dagger U b,$$

and $V, U$, are orthogonal (don’t change size) while we’ve set larger elements of $\hat{\Sigma}^\dagger$ to zero.

- What $p$ to choose? Solution depends sensitively on this since no gap in singular values.
Tikhonov Regularization

- **Idea 2:** add to
  \[ \min_x \| Ax - b \|^2 \]
  a penalty term to get
  \[ \min_x \| Ax - b \|^2 + \lambda^2 R(x). \]

- Most commonly \( R(x) = \| Rx \|^2 \).
  - \( R(\cdot) \) is called regularization constraint
  - \( \lambda \) is called regularization parameter
  - Typical \( R \)'s
    * Identity (size constraint)
    * Gradient, Laplacian, . . . (smoothness constraint)

- **Tradeoff:** Fidelity to data (residual error) for Fidelity to a priori constraint (regularization error)
  - When \( \lambda \) is very small, goes to LS solution
  - When \( \lambda \) is very large, drives solution to 0

- **Key Questions:** How to choose \( R, \lambda \) ??
SVD Analysis of Tikhonov Regularization

- Assume $R = I$, use SVD to analyze Tikhonov regularization
- Then we want to solve
  $$\min_x \| Ax - b \|^2_2 + \lambda^2 \| Ix \|^2_2.$$  
- Note $A^T A = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T$.
- From projection theorem (or algebra) we find that the solution solves:
  $$(A^T A + \lambda^2 I)x = A^T b,$$
  $$(V \Sigma^2 V^T + \lambda^2 VI V^T)x = A^T b,$$
  $$V(\Sigma^2 + \lambda^2 I)V^Tx = A^T b.$$  

where we use the orthogonality of $V$ in the second step.
Continuing the Analysis

- Repeating, we have
  \[ V(\Sigma^2 + \lambda^2 I)V^T x = A^T b \]

- Since \( A^T = V\Sigma U^T \), if we multiply by inverses of everything on the left-hand side and again use orthogonality, we get
  \[ \hat{x} = V(\Sigma^2 + \lambda^2 I)^{-1} \Sigma U^T b. \]

- Compare this to the LS solution
  \[ \hat{x} = V\Sigma^{-2} \Sigma U^T b. \]

- It seems the difference is adding a term to \( \Sigma \) before inverting. Note that this term makes the small diagonal elements larger before we invert !!

- Geometric interpretation: make the small semi-axes larger before inverting . . .

- Idea 1 interpretation: new, safer problem
Analysis in Terms of Singular Values and $\lambda$

- First, note that since both $\Sigma$ and $I$ are diagonal, this is also true for
  1. their sum, and
  2. the inverse of their sum.

In particular that inverse just has the reciprocals of the diagonal elements on it. So $i$’th element of $(\Sigma^2 + \lambda^2 I)^{-1}\Sigma$ is $(\sigma_i)(1/(\sigma_i^2 + \lambda^2))$:

$$\hat{x} = V(\Sigma^2 + \lambda^2 I)^{-1}\Sigma U^T b$$

$$= \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \lambda^2} v_i u_i^T b$$

$$= \sum_{i=1}^{n} \frac{1}{\sigma_i^2 + \lambda^2} v_i u_i^T b$$

$$= \sum_{i=1}^{n} \left( \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \right) \frac{1}{\sigma_i} v_i u_i^T b$$

- Compare this to LS:

$$\hat{x} = \sum_{i=1}^{n} \frac{1}{\sigma_i} v_i u_i^T b.$$
Filter Factors

- The terms that change the result, $\frac{\sigma^2_i}{\sigma^2_i + \lambda^2}$, are called filter factors.
  - For $\sigma_i \gg \lambda$ problem unperturbed
  - For $\sigma_i \ll \lambda$ term goes to 0

So $\lambda$ stabilizes the unregularized solution by filtering the singular values.

- We can reinterpret TSVD as choosing $\lambda$ to be either
  1. 0 (when we keep the singular value) or
  2. $\infty$ (when we throw it away)

- Provides a very nice formalism to understand how regularization works.
- Provides insight into relative size of various choices of $\lambda$. 
Some Other Topics

- General regularization with $R \neq I$: similar analysis using something called the Generalized SVD (GSVD).
- Multiple regularization constraints: topic of current research
- How to choose regularization parameters: coming next, in brief.
- Some other approaches to ill-posed inverse problems.
How to Choose Regularization Parameter?

• Have prior knowledge about noise and signal? Choose $\lambda$ such that

$$\| Ax_\lambda - b \|_2 = \| n \|_2$$

Effectively $\lambda = 1/$SNR.

• No prior knowledge: need a posteriori technique:
  – Generalized Cross Validation
  – L-Curve
  – many others . . .
L-Curve Idea:

- As regularization parameter gets small, solution goes to LS
- As regularization parameter gets big, solution goes to zero
- Plot both errors (residual and regularization penalty) on a log-log scale and choose the good tradeoff point.

Figure from P.C. Hansen
Other Approaches

- Truncated iterative schemes, eg conjugate gradient
- Modified SVD
- Total Variation Approaches
- Frequency Domain, eg Max Entropy and Min Relative Entropy
- Multiple Regularizers
- Admissible Solutions via Convex Optimization
- Parameterize the problem (eg look for spheres or ellipses in DOT)
- Statistical approaches to regularization: generally uses Bayes Rule.