Math 4/896: Seminar in Mathematics Topic: Inverse Theory

Instructor: Thomas Shores
Department of Mathematics

Lecture 28, April 27, 2006 AvH 10

Chapter 5: Tikhonov Regularization

5.2: SVD Implementation of Tikhonov Regularization

SVD Implementation

To solve the Tikhonov regularized problem, first recall:

- $\nabla\left(\|G\mathbf{m}-\mathbf{d}\|_2^2+\alpha^2\|\mathbf{m}\|_2^2\right)=\left(G^TG\mathbf{m}-G^T\mathbf{d}\right)+\alpha^2\mathbf{m}$
- Equate to zero and these are the normal equations for the system $\begin{bmatrix} G \\ \alpha I \end{bmatrix}\mathbf{m}=\begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix}$, or $\left(G^TG+\alpha^2I\right)\mathbf{m}=G^T\mathbf{d}$
- To solve, calculate $\left(G^TG+\alpha^2I\right)^{-1}G^T=V\begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\alpha^2} & & & & \\ & \ddots & & & \\ & & \frac{\sigma_p}{\sigma_p^2+\alpha^2} & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}U^T$

]

SVD Implementation

From the previous equation we obtain that the Moore-Penrose inverse and solution to the regularized problem are given by

$$G_\alpha^\dagger=\sum_{j=1}^{p}\frac{\sigma_j}{\sigma_j^2+\alpha^2}\mathbf{V}_j\mathbf{U}_j^T$$

$$\mathbf{m}_\alpha = G^\dagger \mathbf{d} = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2 + \alpha^2} \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j$$

which specializes to the generalized inverse solution we have seen in the case that G is full column rank and $\alpha = 0$. (Remember $\mathbf{d} = U\mathbf{h}$ so that $\mathbf{h} = U^T\mathbf{d}$.)

]

The Filter Idea

About Filtering:
The idea is simply to "filter" the singular values of our problem so that (hopefully) only "good" ones are used.

- We replace the σ_i by $f(\sigma_i)$. The function f is called a **filter**.
- $f(\sigma) = 1$ simply uses the original singular values.
- $f(\sigma) = \dfrac{\sigma^2}{\sigma^2 + \alpha^2}$ is the Tikhonov filter we have just developed.
- $f(\sigma) = \max\{\text{sgn}(\sigma - \epsilon), 0\}$ is the TSVD filter with singular values smaller than ϵ truncated to zero.

]

The L-curve

L-curves are one tool for choosing the regularization paramter α:

- Make a plot of the curve $(\|\mathbf{m}_\alpha\|_2, \|G\mathbf{m}_\alpha - \mathbf{d}\|_2)$
- Typically, this curve looks to be asymptotic to the axes.
- Choose the value of α closest to the corner.
- Caution: L-curves are NOT guaranteed to work as a regularization strategy.
- An alternative: (Morozov's discrepancy principle) Choose α so that the misfit $\|G\mathbf{m}_\alpha - \mathbf{d}\|_2$ is the same size as the data noise $\|\delta\mathbf{d}\|_2$.

Historical Notes

Tikhonov's original interest was in operator equations

$$d(s) = \int_a^b k(s,t)\, m(t)\, dt$$

or $d = Km$ where K is a compact (**bounded** = **continuous**) linear operator from one Hilbert space H_1 into another H_2. In this situation:

- Such an operator $K : H_1 \to H_2$ has an **adjoint operator** $K^* : H_2 \to H_1$ (analogous to transpose of matrix operator.)
- Least squares solutions to $\min \|Km - d\|$ are just solutions to the **normal** equation $K^*Km = K^*d$ (and exist.)
- There is a Moore-Penrose inverse operator $K^\dagger$ such that $m = K^\dagger d$ is the least squares solution of least 2-norm. But this operator is generally **unbounded** (not continuous.)

Historical Notes

More on Tikhonov's operator equation:

- The operator $(K^*K + \alpha I)$ is bounded with bounded inverse and the **regularized problem** $(K^*K + \alpha I)\, m = K^*d$ has a unique solution m_α.
- Given that $\delta = \|\delta d\|$ is the noise level and that the problem actually solved is $(K^*K + \alpha I)\, m = K^*d^\delta$ with $d^\delta = d + \delta d$ yielding m_α^δ Tikhonov defines a **regular algorithm** to be a choice $\alpha = \alpha(\delta)$ such that

$$\alpha(\delta) \to 0 \text{ and } m^\delta_{\alpha(\delta)} \to K^\dagger d \text{ as } \delta \to 0.$$

- Morozov's discrepancy principle is a regular algorithm.

Finish Section 5.2 by exploring the Example 5.1 file, which constructs the L-curve of the Shaw problem using tools from the Regularization Toolbox.

5.3: Resolution, Bias and Uncertainty in the Tikhonov Solution

]

Resolution Matrix

Definition:
Resolution matrix for a regularized problem starts with this observation:

- Let $G^\natural \equiv \left(G^TG + \alpha^2 I\right)^{-1} G^T$ (generalized inverse)
- Then $\mathbf{m}_\alpha = G^\natural \mathbf{d} = \sum_{j=1}^{p} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\sigma_j} \mathbf{V}_j = VFS^\dagger U^T \mathbf{d}$.
- Model resolution matrix: $R_{\mathbf{m},\alpha} = G^\natural G = VFV^T$
- Data resolution matrix: $R_{\mathbf{d},\alpha} = GG^\natural = UFU^T$

The Example 5.1 file constructs the model resolution matrix of the Shaw problem and shows poor resolution in this case.

5.4: Higher Order Tikhonov Regularization

]

Higher Order Regularization

Basic Idea
We can think of the regularization term $\alpha^2 \|\mathbf{m}\|_2^2$ as favoring minimizing the 0-th order derivative of a function $m(x)$ under the hood. Alternatives:

- Minimize a matrix approximation to $m'(x)$. This is a first order method.
- Minimize a matrix approximation to $m''(x)$. This is a second order method.
- These lead to new minimization problems: to minimize

$$\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2.$$

- How do we resolve this problem as we did with $L = I$?

Example Matrices

We will explore approximations to first and second derivatives at the board.

Key Idea: Generalized SVD (GSVD)

Theorem 1. *Let G be an $m \times n$ matrix and L a $p \times n$ matrix. Then there exist $m \times m$ orthogonal U, $p \times p$ orthogonal V and $n \times n$ nonsingular matrix X with $m \geq n \geq \min\{p, n\} = q$ such that*

$$\begin{aligned} U^T G X &= \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} = \Lambda = \Lambda_{m,n} \\ V^T L X &= \operatorname{diag}\{\mu_1, \mu_2, \ldots, \mu_q\} = M = M_{p,n} \\ \Lambda^T \Lambda + M^T M &= 1. \end{aligned}$$

Also $0 \leq \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n \leq 1$ and $1 \geq \mu_1 \geq \mu_2 \cdots \geq \mu_q \geq 0$.

*The numbers $\gamma_i = \lambda_i / \mu_i$, $i = 1, \ldots, \operatorname{rank}(L) \equiv r$ are called the **generalized singular values** of G and L and $0 \leq \gamma_1 \leq \gamma_2 \cdots \leq \gamma_r$.*

Application to Higher Order Regularization

The minimization problem is equivalent to the problem

$$\left(G^T G + \alpha^2 L^T L\right) \mathbf{m} = G^T \mathbf{d}$$

which has solution forms

$$\mathbf{m}_{\alpha,L} = \sum_{j=1}^{p} \frac{\gamma_j^2}{\gamma_j^2 + \alpha^2} \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\lambda_j} \mathbf{X}_j + \sum_{j=p+1}^{n} \left(\mathbf{U}_j^T \mathbf{d}\right) \mathbf{X}_j$$

Filter factors: $f_j = \dfrac{\gamma_j^2}{\gamma_j^2 + \alpha^2}$, $j = 1, \ldots, p$, $f_j = 1$, $j = p+1, \ldots, n$. Thus

$$\mathbf{m}_{\alpha,L} = \sum_{j=1}^{n} f_j \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{\lambda_j} \mathbf{X}_j.$$

]

Vertical Seismic Profiling Example

The Experiment:
Place sensors at vertical depths z_j, $j = 1, \ldots, n$, in a borehole, then:

- Generate a seizmic wave at ground level, $t = 0$.
- Measure arrival times $d_j = t\left(z_j\right)$, $j = 1, \ldots, n$.
- Now try to recover the slowness function $s\left(z\right)$, given

$$t\left(z\right) = \int_0^z s\left(\xi\right) d\xi = \int_0^\infty s\left(\xi\right) H\left(z - \xi\right) d\xi$$

- It should be easy: $s\left(z\right) = t'\left(z\right)$.
- Hmmm.....or is it?

Do Example 5.4-5.5 from the CD.
]

Model Resolution

Model Resolution Matrix:
As usual, $R_{\mathbf{m},\alpha,L} = G^\natural G$.

- Comment 1.
- Comment 2.
- Comment 3.

TGSVD and GCV

TGSVD:
We have seen this idea before. Simply apply it to formula above, remembering that the generalized singular values are reverse ordered.

- Formula becomes

$$\mathbf{m}_{\alpha,L} = \sum_{j=k}^{p} \frac{\gamma_j^2}{\gamma_j^2 + \alpha^2} \frac{\left(\mathbf{U}_j^T \mathbf{d}\right)}{c_j} \mathbf{X}_j + \sum_{j=p+1}^{n} \left(\mathbf{U}_j^T \mathbf{d}\right) \mathbf{X}_j$$

- Key question: where to start k.

]

GCV

Basic Idea:
Comes from statistical "leave-one-out" cross validation.

- Leave out one data point and use model to predict it.
- Sum these up and choose regularization parameter α that minimizes the sum of the squares of the predictive errors

$$V_0(\alpha) = \frac{1}{m} \sum_{k=1}^{m} \left(\left(G m_{\alpha,L}^{[k]} \right)_k - d_k \right)^2 .$$

- One can show a good approximation is

$$V_0(\alpha) = \frac{m \left\| G\mathbf{m}_\alpha - \mathbf{d} \right\|_2}{\mathrm{Tr}\left(I - GG^\natural\right)^2}$$

Example 5.6-7 gives a nice illustration of the ideas. Use the CD script to explore it. Change the startupfile path to Examples/chap5/examp6, then examp7.

Error Bounds

]

Error Bounds

Error Estimates:
They exist, even in the hard cases where there is error in both G and d.

- In the simpler case, G known exactly, they take the form

$$\frac{\|\mathbf{m}_\alpha - \widetilde{\mathbf{m}}_\alpha\|_2}{\|\mathbf{m}_\alpha\|_2} \leq \kappa_\alpha \frac{\left\|\mathbf{d} - \widetilde{\mathbf{d}}\right\|_2}{\|G\mathbf{m}_\alpha\|_2}$$

  where κ_α is inversely proportional to α.

]

Error Bounds

More Estimates:

- Suppose that the true model $\mathbf{m}_{true}$ is "smooth" in the sense that there exists vector $\mathbf{w}$ such that $(p = 1)$ $\mathbf{m}_{true} = G^T\mathbf{w}$ or $(p = 2)$ $\mathbf{m}_{true} = G^TG\mathbf{w}$. Let $\Delta = \delta/\|\mathbf{w}\|$ and $\gamma = 1$ if $p = 1$ and $\gamma = 4$ if $p = 2$. Then the choice $\widehat{\alpha} = (\Delta/\gamma)^{1/(p+1)}$ is optimal in the sense that we have the error bound

$$\left\|\mathbf{m}_{true} - G^\natural\mathbf{d}\right\|_2 = \gamma\,(p+1)\,\widehat{\alpha}^p = \mathcal{O}\left(\Delta^{\frac{p}{p+1}}\right).$$

- This is about the best we can do. Its significance: the best we can hope for is about 1/2 or 2/3 of the significant digits in the data.

Chapter 6: Iterative Methods – A Brief Discussion

]

Image Recovery

Problem:
An image is blurred and we want to sharpen it. Let intensity function $I_{true}\,(x, y)$define the true image and $I_{blurred}\,(x, y)$ define the blurred image.

- A typical model results from convolving true image with Gaussian point spread function

$$I_{blurred}\,(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_{true}\,(x - u, y - v)\,\Psi\,(u, v)\;du\,dv$$

  where $\Psi\,(u, v) = e^{-\left(u^2+v^2\right)/\left(2\sigma^2\right)}$.

- Think about discretizing this over an SVGA image (1024×768).
- But the discretized matrix should be sparse!

]

Sparse Matrices and Iterative Methods

Sparse Matrix:
A matrix with sufficiently many zeros that we should pay attention to them.

- There are efficient ways of storing such matrices and doing linear algebra on them.
- Given a problem $A\mathbf{x} = \mathbf{b}$ with A sparse, iterative methods become attractive because they usually only require storage of A, $\mathbf{x}$ and some auxillary vectors, and saxpy, gaxpy, dot algorithms – ("scalar a*x+y" , "general A*x+y", "dot product")
- Classical methods: Jacobi, Gauss-Seidel, Gauss-Seidel SOR and conjugate gradient.
- Methods especially useful for tomographic problems: Kaczmarz's method, ART (algebraic reconstruction technique).

]

Yet Another Regularization Idea

To regularize in face of iteration:
Use the number of iteration steps taken as a regularization parameter.

- Conjugate gradient methods are designed to work with SPD coefficient matrices A in the equation $A\mathbf{x} = \mathbf{b}$.
- So in the unregularized least squares problem $G^T G\mathbf{m} = G^T\mathbf{d}$ take $A = G^T G$ and $\mathbf{b} = G^T\mathbf{d}$, resulting in the CGLS method, in which we avoid explicitly computing $G^T G$.
- Key fact: in exact arithmetic, if we start at $\mathbf{m}^{(0)} = \mathbf{0}$, then $\left\|\mathbf{m}^{(k)}\right\|$ is monotone increasing in k and $\left\|G\mathbf{m}^{(k)} - \mathbf{d}\right\|$ is monotonically decreasing in k. So we can make an L-curve in terms of k.

Do Example 6.3 from the CD. Change startupfile path to Examples/chap6/examp3

Chapter 7: Additional Regularization Techniques

7.1: Using Bounds as Constraints

]

Regularization...Sort Of

Basic Idea:
Use prior knowledge about the nature of the solution to restrict it:

- Most common restrictions: on the magnitude of the parameter values. Which leads to the problem:
- Minimize $f(\mathbf{m})$ subject to $l \leq m \leq u$.
- One could choose $f(\mathbf{m}) = \|G\mathbf{m} - \mathbf{d}\|_2$(BVLS)
- One could choose $f(\mathbf{m}) = \mathbf{c}^T \cdot \mathbf{m}$with additional constraint $\|G\mathbf{m} - \mathbf{d}\|_2 \leq \delta$.

Example 3.3

Contaminant Transport
Let $C(x,t)$ be the concentration of a pollutant at point x in a linear stream, time t, where $0 \leq x < \infty$ and $0 \leq t \leq T$. The defining model

$$\begin{aligned} \frac{\partial C}{\partial t} &= D\frac{\partial^2 C}{\partial x^2} - v\frac{\partial C}{\partial x} \\ C(0,t) &= C_{in}(t) \\ C(x,t) &\rightarrow 0,\ x \rightarrow \infty \\ C(x,0) &= C_0(x) \end{aligned}$$

Solution

Solution:
In the case that $C_0(x) \equiv 0$, the explicit solution is

$$C(x,T) = \int_0^T C_{in}(t)\, f(x, T-t)\, dt,$$

where

$$f(x,\tau) = \frac{x}{2\sqrt{\pi D \tau^3}} e^{-(x-v\tau)^2/(4D\tau)}$$

Inverse Problem

Problem:
Given simultaneous measurements at time T, to estimate the contaminant inflow history. That is, given data

$$d_i = C\left(x_i, T\right),\ i = 1, 2, \ldots, m,$$

to estimate

$$C_{in}\left(t\right),\ 0 \leq t \leq T.$$

Change the startupfile path to Examples/chap7/examp1 execute it and examp.

7.2: Maximum Entropy Regularization

]

A Better Idea (?)

Entropy:
$E\left(\mathbf{m}\right) = -\sum_{j=1}^{n} m_j \ln\left(w_j m_j\right)$, $\mathbf{w}$ a vector of positive weights.

- Motivated by Shannon's information theory and Bolzmann's theory of entropy in statistical mechanics. A measure of uncertainty about which message or physical state will occur.
- Shannon's entropy function for a probability distribution $\{p_i\}_{i=1}^{n}$ is $H\left(\mathbf{p}\right) = -\sum_{i=1}^{n} p_i \ln\left(p_i\right)$.
- Bayesian Maximimum Entropy Principle: least biased model is one that maximizes entropy subject to constraints of testable information like bounds or average values of parameters.

Maximum Entropy Regularization

Maximize Entropy:
That is, our version. So problem looks like:

- Maximize $-\sum_{j=1}^{n} m_j \ln\left(w_j m_j\right)$
- Subject to $\left\|G\mathbf{m} - \mathbf{d}\right\|_2 \leq \delta$ and $\mathbf{m} \geq \mathbf{0}$.

- In absence of extra information, take $w_i = 1$. Lagrange multipliers give:
- Minimize $\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha^2 \sum_{j=1}^{n} m_j \ln(w_j m_j)$,
- subject to $\mathbf{m} \geq \mathbf{0}$.

Change the startupfile path to Examples/chap7/examp2 execute it and examp.

7.3: Total Variation

]

TV Regularization

We only consider total variation regularization from this section.

Regularization term:

$\mathrm{DV}(\mathbf{m}) = \sum_{j=1}^{n-1} |m_{j+1} - m_j| = \|L\mathbf{m}\|_1$, where L is the matrix used in first order Tikhonov regularization.

- Problem becomes: minimize $\|G\mathbf{m} - \mathbf{d}\|_2^2 + \alpha \|\mathbf{m}\|_1$
- Better yet: minimize $\|G\mathbf{m} - \mathbf{d}\|_1 + \alpha \|\mathbf{m}\|_1$.
- Equivalently: minimize $\left\| \begin{bmatrix} G \\ \alpha L \end{bmatrix} \mathbf{m} - \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \right\|_1$.
- Now just use IRLS (iteratively reweighted least squares) to solve it and an L-curve of sorts to find optimal α.
-

]

Total Variation

Key Property:

- TV doesn't smooth discontinuities as much as Tikhonov regularization.

Change startupfile path to Examples/chap7/examp3 execute it and examp.

Chapter 9: Nonlinear Regression

Newton's Method

]

Basic Problems

Root Finding:
Solve the system of equations represented in vector form as

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}.$$

for point(s) $\mathbf{x}^*$ for which $\mathbf{F}(\mathbf{x}^*) = \mathbf{0}$.

- Here $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ and $x = (x_1, \ldots, x_m)$
- Gradient notation: $\nabla f_j(\mathbf{x}) = \left(\frac{\partial f_j}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f_j}{\partial x_m}(\mathbf{x})\right)$.
- Jacobian notation: $\nabla \mathbf{F}(\mathbf{x}) = [\nabla f_1(\mathbf{x}), \ldots, \nabla f_m(\mathbf{x})]^T = \left[\frac{\partial f_i}{\partial x_j}\right]_{i,j=1,\ldots m}$.

]

Basic Problems

Optimization:
Find the minimum value of scalar valued function $f(\mathbf{x})$, where $\mathbf{x}$ ranges over a feasible set Ω.

- Set $\mathbf{F}(\mathbf{x}) = \nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_m}(\mathbf{x})\right)$
- Hessian of f: $\nabla(\nabla f(\mathbf{x})) \equiv \nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]$.

]

Taylor Theorems

First Order
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has continuous second partials and $\mathbf{x}^*, \mathbf{x} \in \mathbb{R}^n$. Then $f(\mathbf{x}) = f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) + \mathcal{O}\left(\|\mathbf{x} - \mathbf{x}^*\|^2\right), \mathbf{x} \to \mathbf{x}$.

Second Order

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has continuous third partials and $\mathbf{x}^*, \mathbf{x} \in \mathbb{R}^n$. Then $f(\mathbf{x}) = f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \nabla^2 f(\mathbf{x}^*) (\mathbf{x} - \mathbf{x}^*) + \mathcal{O}\left(\|\mathbf{x} - \mathbf{x}^*\|^3\right), \mathbf{x} \to \mathbf{x}$.

(See Appendix C for versions of Taylor's theorem with weaker hypotheses.)

]

Newton Algorithms

Root Finding

Input $\mathbf{F}$, $\nabla\mathbf{F}$, $\mathbf{x}^0$, N_{max}
  for $k = 0, ..., N_{max}$
    $\mathbf{x}^{k+1} = \mathbf{x}^k - \nabla\mathbf{F}\left(\mathbf{x}^k\right)^{-1} \mathbf{F}\left(\mathbf{x}^k\right)$
    if $\mathbf{x}^{k+1}, \mathbf{x}^k$ pass a convergence test
      return$(\mathbf{x}^k)$
    end
  end
  return$(\mathbf{x}^{N_{max}})$

Convergence Result

Theorem 2. *Let $\mathbf{x}^*$ be a root of the equation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$, where $\mathbf{F}, \mathbf{x}$ are m-vectors, $\mathbf{F}$ has continuous first partials in some neighborhood of $\mathbf{x}^*$ and $\nabla\mathbf{F}(\mathbf{x}^*)$ is non-singular. Then Newton's method yields a sequence of vectors that converges to $\mathbf{x}^*$, provided that $\mathbf{x}^0$ is sufficiently close to $\mathbf{x}^*$. If, in addition, $\mathbf{F}$ has continuous second partials in some neighborhood of $\mathbf{x}^*$, then the convergence is quadratic in the sense that for some constant $K > 0$,*

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leq K \left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2.$$

]

Newton for Optimization

Bright Idea:
We know from calculus that where $f(\mathbf{x})$ has a local minimum, $\nabla f = \mathbf{0}$. So just let $\mathbf{F}(\mathbf{x}) = \nabla f(\mathbf{x})$ and use Newton's method.

- Result is iteration formula: $\mathbf{x}^{k+1} = \mathbf{x}^k - \nabla^2 f\left(\mathbf{x}^k\right)^{-1} \nabla f\left(\mathbf{x}^k\right)$
- We can turn this approach on its head: root finding is just a special case of optimization, i.e., solving $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ is the same as minimizing $f(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\|^2$.

- Downside of root finding point of view of optimization: saddle points and local maxima $\mathbf{x}$ also satisfy $\nabla f(\mathbf{x}) = \mathbf{0}$.
- Upside of optimization view of root finding: if $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ doesn't have a root, minimizing $f(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\|^2$ finds the next best solutions – least squares solutions!
- In fact, least squares problem for$\|G\mathbf{m} - \mathbf{d}\|^2$ is optimization!

Remarks on Newton

About Newton:
This barely scratches the surface of optimization theory (take Math 4/833 if you can!!).

- Far from a zero, Newton does not exhibit quadratic convergence. It is accelerated by a line search in the Newton direction $-\nabla\mathbf{F}\left(\mathbf{x}^k\right)^{-1}\mathbf{F}\left(\mathbf{x}^k\right)$ for a point that (approximately) minimizes a merit function like $m(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\|^2$.
- Optimization is NOT a special case of root finding. There are special characteristics of the min $f(\mathbf{x})$ problem that get lost if one only tries to find a zero of ∇f .
- For example, $-\nabla f$ is a search direction that leads to the method of steepest descent. This is not terribly efficient, but well understood.
- There is an automatic merit function, namely $f(\mathbf{x})$, in any search direction. Using this helps avoid saddle points, maxima.

Gauss-Newton and Levenberg-Marquardt Methods

Gauss-Newton and Levenberg-Marquardt

The Problem:
Given a function $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, minimize $f(\mathbf{x}) = \sum_{k=0}^{m} f_k(\mathbf{x})^2 = \|\mathbf{F}(\mathbf{x})\|^2$.

- Newton's method can be very expensive, due to derivative evaluations.
- For starters, one shows $\nabla f(\mathbf{x}) = 2(\nabla\mathbf{F}(\mathbf{x}))^T\mathbf{F}(\mathbf{x})$
- Then, $\nabla^2 f(\mathbf{x}) = 2(\nabla\mathbf{F}(\mathbf{x}))^T\nabla\mathbf{F}(\mathbf{x}) + Q(\mathbf{x})$, where $Q(\mathbf{x}) = \sum_{k=1}^{m} f_k(\mathbf{x})\nabla^2 f_k(\mathbf{x})$ contains all the second derivatives.

LM

The Problem:
Given a function $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, minimize $f(\mathbf{x}) = \sum_{k=0}^{m} f_k(\mathbf{x})^2 = \|\mathbf{F}(\mathbf{x})\|^2$.

- This inspires a so-called quasi-Newton method, which approximates the Hessian as $\nabla^2 f(\mathbf{x}) \approx 2(\nabla \mathbf{F}(\mathbf{x}))^T \nabla \mathbf{F}(\mathbf{x})$.
- Thus, Newton's method morphs into the Gauss-Newton (GN) method

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \left(\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^{-1} \left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \mathbf{F}\left(\mathbf{x}^k\right)$$

- There's a problem here. See it?

LM

The Problem:
$\nabla \mathbf{F}(\mathbf{x})$ may not have full column rank.

- A remedy: regularize the Newton problem to $\left(\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \nabla \mathbf{F}\left(\mathbf{x}^k\right) + \lambda_k I \right) \mathbf{p} = -\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \mathbf{F}\left(\mathbf{x}^k\right)$ with λ suitably chosen positive number for $\mathbf{p} = \mathbf{x} - \mathbf{x}^k$
- In fact, Lagrange multipliers show we are really solving a constrained problem of minimizing $\left\| \nabla \mathbf{F}\left(\mathbf{x}^k\right) \mathbf{p} + \mathbf{F}\left(\mathbf{x}^k\right) \right\|^2$ subject to a constraint $\|\mathbf{p}\| \leq \delta_k$. Of course, δ_k determines λ_k and vice-versa.
- The idea is to choose λ_k at each step: Increase it if the reduction in $f(\mathbf{x})$ was not as good as expected, and decrease it if the reduction was better than expected. Otherwise, leave it alone.

LM

More on LM:
$\left(\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \nabla \mathbf{F}\left(\mathbf{x}^k\right) + \lambda_k I \right) \mathbf{p} = -\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \mathbf{F}\left(\mathbf{x}^k\right)$.

- For small λ_k, LM becomes approximately $\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \nabla \mathbf{F}\left(\mathbf{x}^k\right) \mathbf{p} = -\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \mathbf{F}\left(\mathbf{x}^k\right)$ which is GN with its favorable convergence rate.
- For large λ_k, LM becomes approximately $\mathbf{p} = -\frac{1}{\lambda_k} \left(\nabla \mathbf{F}\left(\mathbf{x}^k\right) \right)^T \mathbf{F}\left(\mathbf{x}^k\right)$, which is a steepest-descent step, slow but convergent.

- For large λ_k, LM becomes approximately $\mathbf{p} = -\frac{1}{\lambda_k}\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right)\right)^T \mathbf{F}\left(\mathbf{x}^k\right)$, which is a steepest-descent step, slow but convergent.
- For small residuals, LM (and GN, when stable) converge superlinearly. They tend to perform poorly on large residual problems, where the dropped Hessian terms are significant.

LM

Another Perspective on LM:

- NB: λ_k is not a regularizaton parameter in usual sense, but rather a tool for efficiently solving a nonlinear system which itself may or may not be regularized.
- However: suppose our objective is to find a least squares solution to the problem $\mathbf{F}(\mathbf{x}) = \mathbf{d}$, given output data $\mathbf{d}$ with error, in the form of $\mathbf{d}^\delta$, i.e., to minimize $\left\|\mathbf{F}(\mathbf{x}) - \mathbf{d}^\delta\right\|^2$.
- In this case, LM amounts to cycles of these three steps:
- Forward-solve: compute $\mathbf{d}^k = \mathbf{F}\left(\mathbf{x}^k\right)$.
- Linearize: $\nabla \mathbf{F}\left(\mathbf{x}^k\right)\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right) = \mathbf{d}^\delta - \mathbf{d}^k$.
- Regularize: $\left(\left(\nabla \mathbf{F}\left(\mathbf{x}^k\right)\right)^T \nabla \mathbf{F}\left(\mathbf{x}^k\right) + \alpha_k I\right)\mathbf{p} = \left(\nabla \mathbf{F}\left(\mathbf{x}^k\right)\right)^T \left(\mathbf{d}^\delta - \mathbf{d}^k\right)$
- This is a regularization technique for nonlinear problems and is called **output least squares**.

Section 9.3: Statistical Aspects

Statistics

Problem is $G(\mathbf{m}) = \mathbf{d}$ with least squares solution $\mathbf{m}^*$:
Now what? What statistics can we bring to bear on the problem?

- We minimize $\|\mathbf{F}(\mathbf{m})\|^2 = \sum_{i=1}^{n} \frac{\left(G(\mathbf{m}) - d_i\right)^2}{\sigma_i^2}$
- Treat the linear model as locally accurate, so misfit is $\nabla \mathbf{F} = \mathbf{F}(\mathbf{m} + \Delta \mathbf{m}) - \mathbf{F}(\mathbf{m}^*) \approx \nabla \mathbf{F}(\mathbf{m}^*)\nabla \mathbf{m}$
- Obtain covariance matrix $\mathrm{Cov}(\mathbf{m}^*) = \left(\nabla \mathbf{F}(\mathbf{m}^*)^T \nabla \mathbf{F}(\mathbf{m}^*)\right)^{-1}$

- If σ is unknown but constant across measurements, take $\sigma_i = 1$ above and use for σ in $\frac{1}{\sigma^2}\left(\nabla \mathbf{F}\left(\mathbf{m}^*\right)^T \nabla \mathbf{F}\left(\mathbf{m}^*\right)\right)^{-1}$ the estimate

$$s^2 = \frac{1}{m-n}\sum_{i=1}^{m}\left(G\left(\mathbf{m}\right) - d_i\right)^2.$$

- Do confidence intervals, χ^2 statistic and p-value as in Chapter 2.

Implementation Issues

Implementation Issues

What could go wrong?

- Problem may have many local minima.
- Even if it has a unique solution, it might lie in a long flat basin.
- Analytical derivatives may not be available. This presents an interesting regularization issue not discussed by the authors. We do so at the board.
- One remedy for first problem: use many starting points and statistics to choose best local minimum.
- One remedy for second problem: use a better technique than GN or LM.
- Do Example 9.2 from the CD to illustrate some of these ideas.
- If time permits, do data fiting from Great Britian population data.

Chapter 10: Nonlinear Inverse Problems

Regularizing Nonlinear Least Squares Problems

Penalized (Damped) Least Squares

Basic Problem:
Solve $G\left(\mathbf{m}\right) = \mathbf{d}$, where G is a nonlinear function. As usual, $\mathbf{d}$ will have error and this may not be a well-posed problem. Assume variables are scaled, so standard deviations of measurements are incorporated. So we follow the same paths as in Chapter 5.

- Recast: minimize $\|\mathbf{Gm} - \mathbf{d}\|_\mathbf{2}$ – unconstrained least squares.
- Recast: minimize $\|G\mathbf{m} - \mathbf{d}\|_2$ subject to $\|L\mathbf{m}\|_2 \leq \epsilon$, where L is a damping matrix (e.g., $L = I$.)

- Recast: minimize $\|L\mathbf{m}\|_2$ subject to $\|\mathbf{Gm}-\mathbf{d}\|_\mathbf{2} \leq \delta$.
- Recast: (**damped least squares**) minimize $\|G\mathbf{m}-\mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2$. This is also a **Tikhonov regularization** of the original problem, possibly higher order.
- Method of Lagrange multipliers doesn't care if G is nonlinear, so we can apply it as in Chapter 5 to show that these problems are essentially equivalent.
- A big difference is that we can no longer derive linear normal equations for the least squares problem.

Solution Methodology: Penalized Least Squares

Basic Idea:
Regularize, then linearize.

- Regularize: $\|G(\mathbf{m})-\mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2$.
- Equivalently: minimize $\left\| \begin{bmatrix} G(\mathbf{m})-\mathbf{d} \\ \alpha L\mathbf{m} \end{bmatrix} \right\|_2^2 \equiv \|H(\mathbf{m})\|_2^2$.
- Linearize: Compute the Jacobian of this vector function: $\nabla H(\mathbf{m}) = \begin{bmatrix} \nabla G(\mathbf{m}) \\ \alpha L \end{bmatrix}$.
- The linear model of G near current guesstimate $\mathbf{m}^k$, with $\Delta\mathbf{m} = \mathbf{m}-\mathbf{m}^k$: $G(\mathbf{m}) \approx G(\mathbf{m}) + \nabla G(\mathbf{m})\Delta\mathbf{m}$.
- This leads to the system

$$\left(\nabla G\left(\mathbf{m}^k\right)^T \nabla G\left(\mathbf{m}^k\right) + \alpha^2 L^T L\right)\Delta\mathbf{m} = -\nabla G\left(\mathbf{m}^k\right)^T \left(G\left(\mathbf{m}^k\right)-\mathbf{d}\right) - \alpha^2 L^T L\mathbf{m}^k$$

Work through Example 10.1 of CD.

Occam's Inversion

Solution Methodology: An Output Least Squares

Basic Idea:
Linearize, then regularize. Authors call this method "Occam's inversion" – it is a special type of output least squares.

- Develop the linear model of $G(\mathbf{m})$ near $\mathbf{m}^k$: $G(\mathbf{m}) \approx G\left(\mathbf{m}^k\right) + \nabla G\left(\mathbf{m}^k\right)\left(\mathbf{m}-\mathbf{m}^k\right)$

- Linearize $\|G(\mathbf{m}) - \mathbf{d}\|_2^2 + \alpha^2 \|L\mathbf{m}\|_2^2$ by making the above replacement for $G(\mathbf{m})$. Call the solution $\mathbf{m}^{k+1}$.
- This leads to the system $\mathbf{m}^{k+1} = \left(\nabla G\left(\mathbf{m}^k\right)^T \nabla G\left(\mathbf{m}^k\right) + \alpha^2 L^T L\right)^{-1} \nabla G\left(\mathbf{m}^k\right)^T \widehat{\mathbf{d}}\left(\mathbf{m}^k\right)$, where $\widehat{\mathbf{d}}\left(\mathbf{m}^k\right) = d - G\left(\mathbf{m}^k\right) + \nabla G\left(\mathbf{m}^k\right)^T \mathbf{m}^k$.
- The algorithm is to solve this equation with initial guess $\mathbf{m}^0$, but at each iteration choose the largest value of α such that $\chi^2\left(\mathbf{m}^{k+1}\right) \leq \delta^2$. If none, pick value of α that minimizes χ^2. Stop if and when sequence converges to a solution with $\chi^2 \leq \delta^2$.

Examples

Example 10.2:
We are to estimate subsurface electrical conductivity from above ground EM induction measurements. Used is a Geonics EM-38. The model is complex and we treat it as a black box. Since Jacobians are lacking, we simply use finite differences to approximate them. Measurements are taken at heights of $0, 10, 20, 30, 40, 50, 75, 100$ and 150 cm above the surface. subsurface is discretized into 10 layers, each 20 cm thick with a bottom semi-infinite layer. Now modify the path to Examples/chap10/examp2 and run examp.

]

Examples

Population Example:
Refer to the file N2.pdf for background.

- It's convenient to write $P(t) = \dfrac{KP_0}{(K - P_0)\, e^{-rt} + P_0}$.
- Play with some starting points and graph the resulting model against the observed values.
- How good is this? Can we do better? Do we need to regularize? Is the model appropriate? Discuss.

]

Which Brings Us To...

THE END

(not quite...) Final Review

]

Final Review

Rules of the Game:

- The final will have two parts.
- An in-class part that is closed book, closed notes, NO calculators, laptops, cell phones, blackberries, etc., etc. This part is worth 80 points. This exam will be administered on Monday, May 1, 8:30-10:30 pm.
- A take-home part that is worth 50 points. This exam will be available on the class home page Tuesday morning. It will be due exactly 3 days after you received a copy. It can either be hardcopy or in the form of a pdf (or even Microsoft Word) file which you can email to me. You must show all your work, including copies of any scripts that you used and their relevant outputs.

Final Review

More Rules of the Game:

- There is to be absolutely no consultation of any kind with anyone else other than me about the exam. If there are points of clarification or errors, I will post them on our message board.
- ALL materials used in your work that have not been provided by me for this course must be explicitly credited in your write-up.

]

Final Review

Material Covered in Final:

- Lecture notes in Math492s6LecturesPostMid.pdf.
- Homework problems (fairly simple questions or problems.)
- Excludes all lecture notes material covering our intro/review of Matlab, linear algebra and probability/statistics and pre-midterm material *per se*.
- +-> There will be 6 questions.

]

Sample Questions

Sample Questions:

- What is the L-curve? Describe how it is used.
- Discuss the difference between penalized least squares and output least squares.
-

Sample Questions

Sample Questions:

- Explain what second order regularization means. Why is it describes as biased (no proofs)?
- State the GSVD Theorem and one application of it (you do not have to prove it.) Use the statement of the theorem to describe the generalized singular values.