SIOG 231 GEOMAGNETISM AND ELECTROMAGNETISM

Lecture 17: Regularized Inversion

Introduction

With properly migrated seismic reflection data you can see the geology in the patterns of reflections. Similarly, with aeromagnetic data the tectonic fabric can be seen in the patterns of magnetic intensity. But with magnetotelluric and other electromagnetic data, you just have some complex numbers as a function of frequency. At best, you could inspect the apparent resistivity curves and tell whether resistivity increases or decreases with depth. And even that could be misleading in complex geology. We have seen how nonlinear parameter fitting can handle simple problems such as a small number of layers. However, as you increase the number of parameters this approach becomes unstable. Models will either simply not converge, or the parameters will oscillate and become extreme. Indeed, we know from the D+ solution that the true least squares solution in 1D is extreme, and something similar happens in 1D resistivity. We don't have such analytical solutions in 2D and 3D, but we do know that as you try to drive down misfit to the data, the models do become rougher and rougher. The solution to this is to introduce some form of *regularization*, and minimize the roughness in the model as well as the misfit to the data.

We have seen that for some simple problems, least squares **analytical solutions** can be found. Although important from an inverse theory point of view, the models are not geologically useful. The 1D MT sounding problem is one of the few (3?) geophysical problems for which an analytical least squares solution exists, derived by Parker and Whaler (1981) and called the D^+ solution. Unfortunately, the D^+ solution, although best fitting in a least squares sense, is pathologically rough (delta functions of conductance (conductivity times thickness) in an infinitely resistive half-space). However, the misfit measure obtained from D^+ can be a useful guide to data quality.

One approach to fitting models to data is **trial and error modeling.** This can work quite easily for simple 1D models but becomes challenging in 2D (and probably impossible in 3D). However, prior to the development of 2D regularized inversion codes around 1990 this was the only possible way, and Phil Wanamaker produced a complicated 2D model of the EMSLAB data set in 1988, guided no doubt by lots of prior understanding of the geology.

Another approach is **stochastic modeling**, also called Monte Carlo modeling, in which models are generated quasi-randomly and tested to see if they fit the data. Truly random model generation would be prohibitively inefficient, and so this approach relies on algorithms that increase the efficiency of the model generation, which include Markov chains, genetic or evolutionary algorithms, simulated annealing, etc. These approaches are often classed as Beyesian inversion, although strictly speaking most approaches to inversion, including regularized inversion, can be viewed through the lens of Bayes' theorem. Because literally millions of models and forward calculations are required, stochastic inversion is still limited to relatively sparse parameterizations. The advantages are that one can generate some measure of the uncertainty in the model, one should be able to avoid local minima in misfit space, and only forward model calculations are required (no Jacobian matrix is needed).

Deterministic methods use the Jacobian matrix to guide a solution from a starting model to a model that fits the data. We have seen how this works when we discussed the Marquardt method. Other Newton methods are used in higher dimensions, and the conjugate gradient approach only uses the Jacobian matrix a single

column at a time, avoiding the storage and inversion of what can be a very large matrix for 3D problems. For all but sparsely parameterized 1D problems, deterministic approaches are invariably regularized.

For non-linear problems, one cannot generally guarantee that deterministic inversions will converge to the global minimum of misfit. Since they are designed to find a minimum, they can get "stuck" in a local minimum, especially if they are dependent on the starting model.

It is worth noting that all these approaches are exercises in model construction, and all these models are non-unique and uncertain. A true inverse theoretician would seek something that the data could constrain uniquely. A good example of this is excess mass in gravity. For EM methods, this approach is limited to rigorous bounds on average conductivity over some depth range, as was done by Medin et al. (2017).

For most problems, geophysical inversion is **non-unique**, in that if you have a model space (defined by your forward modeling – 1D, 2D, or 3D with a given parametrization) that can fit the data adequately, then an infinite number of models will fit the data. If your model space cannot fit the data (say you are trying to fit 3D MT data using a 2D model) then no models can be found. Thus, a single misfit in misfit space will map into an infinite number of models, or none at all.

Geophysical inversion is also usually poorly constrained, or **ill-posed**, in that a small variation in misfit can map to a large distance in model space. Again, the D+ solution provides some insight here – the last little bit of misfit improvement is associated with an infinite distance in 1D model space, assuming that your model space can accommodate delta functions.

Regularization

We have seen that the true LS solution is too extreme to be useful and sparse parameterizations limit the solution to a fixed, small number of layers. Indeed, the inversion of the curvature matrix in the sparse parameter solution provided by the Marquardt algorithm isn't defined when the model parameters exceed the number of data, and since individual data points are rarely fully independent (i.e. MT sounding curves are smooth functions) this problem arises much sooner than this limit. Also, when a Marquardt inversion is given so many parameters that the individual parameters are not independently resolvable, the solution becomes unstable, either oscillating wildly or simply not converging. Since the model update is solved for, there will inevitably be an imprint of the starting model in any solution that is obtained.

What to do? One approach, suggested by Backus and Gilbert (1967), is to allow a large number of parameters but minimize $\Delta\mathbf{m}$. This and related algorithms converge extremely slowly and are called by Parker (1994) *creeping methods*. Almost all high-dimensional inversion today incorporates some type of *regularization*, an approach suggested by Tikhonov and Arsenin (1977), which explicitly penalizes bad behavior in the model. Instead of must minimizing the misfit to the data, we add a $\mu||\mathbf{Rm}||^2$ term, where $\mathbf{Rm}$ is some measure of model roughness or complexity, and μ is a trade-off parameter:

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2\right) + \mu||\mathbf{Rm}||^2$$

The roughness measure, $\mathbf{Rm}$ is often taken to be first differences between adjacent model parameters and easily generated by a matrix $\mathbf{R}$ consisting of (-1, 1) entries. For a stack of layers $\mathbf{R}$ is a diagonal matrix that looks like:

$$\mathbf{R}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ & & \ddots & & & \ddots & \\ & & & & & & \\ & & & & & -1 & 1 \end{pmatrix}$$

m_1	-1						
m_2	+1	-1					
m_3		+1	-1				
m_4			+1	-1			
m_5				+1	-1		
m_6					+1	-1	
m_7						+1	-1
m_8							+1

For a second difference roughness you could use

$$\mathbf{R}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & -2 & 1 & \dots & 0 \\ & & \ddots & & & \ddots & \\ & & & & & & \\ & & & & 1 & -2 & 1 \end{pmatrix} . \tag{1}$$

We can weight the rows of $\mathbf{R}$ with layer thickness, depth, or whatever we desire. We can even neglect them or zero them out (which amounts to the same thing) if we want to allow an un-penalized jump to appear in the model. When μ is small, the model roughness is ignored and we try and fit the model. Indeed, if $\mu = 0$ then we just have a least squares solution again. When μ is large, the misfit is ignored and we try and we try to reduce the roughness of the model, that is, make it smooth.

So how to choose μ? One approach is to choose a fixed μ based on experience or some algorithm, and then minimize U. However, this is in effect determining *a priori* how rough your model will be, and we don't know much about the model – presumably that is why we collected the data in the first place. On the other hand, if we have done a good job of estimating our error bars, we should have a good idea of how well we should be fitting our data, and we can choose a target misfit, χ^2_* which is greater than the minimum possible, but statistically acceptable. For well behaved, well estimated errors, this will be close the the expectation value, M, or equivalent to RMS = 1. Then we have

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2 - \chi^2_*\right) + \mu||\mathbf{Rm}||^2$$

or equivalently

$$U = ||\mathbf{Rm}||^2 + \mu^{-1}\left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2 - \chi^2_*\right)$$

where μ^{-1} is a Lagrange multiplier, such that we minimize the roughness of the model subject to the constraint that the second term goes to zero (i.e. $||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2 = \chi^2_*$).

For example, instead of minimizing χ^2, we minimize an unconstrained functional

Next we linearize $f(\mathbf{m})$ around a starting model $\mathbf{m}_0$ in the usual way, introducing the Jacobian matrix $\mathbf{J}$

$$U = ||\mathbf{Rm}_1||^2 + \mu^{-1}\left(||\mathbf{Wd} - \mathbf{W}\big(f(\mathbf{m}_0) + \mathbf{J}(\mathbf{m}_1 - \mathbf{m}_0)\big)||^2 - \chi^2_*\right)$$

and noting that we apply the roughness penalty to the new model, not the starting model, and differentiate U with respect to $\mathbf{m}_1$, which may be rearranged to get $\mathbf{m}_1$:

$$\mathbf{m}_1 = \left(\mu\mathbf{R}^T\mathbf{R} + (\mathbf{WJ})^T\mathbf{WJ}\right)^{-1}(\mathbf{WJ})^T\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}) + \mathbf{Jm}_0) \quad .$$

Compare this to the Gauss-Newton iteration:

$$\Delta\mathbf{m} = \left((\mathbf{WJ})^T\mathbf{WJ}\right)^{-1}(\mathbf{WJ})^T\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}))$$

We need only to choose the tradeoff (Lagrange) multiplier μ. The approach of Constable *et al.* (1987) was to note that for each iteration χ^2 is a function of μ, and to use 1D optimization (simply a line search) to minimize χ^2 when $\chi^2 > \chi^2_*$ and to find μ such that $\chi^2 = \chi^2_*$ otherwise. Constable *et al.* called this approach 'Occam's inversion'. Although the Occam algorithm is reliable and has good convergence behavior, the computation and storage of **J** for large models can be limiting, but for 1D and 2D models there is no problem.

There are many advantages to this approach:

We are solving for the next model, $\mathbf{m}_1$, directly, not the model update $\Delta\mathbf{m}$, so there is less memory of the starting model and we can take large steps in model space. Thus convergence is fast.

The matrix inversion is stabilized by the addition of the diagonal matrix $\mathbf{R}^T\mathbf{R}$, not unlike how Marquardt inversion is stabilized.

The smoothing term defines model parameters that are not resolved by the data, again stabilizing the inversion.

The smooth models are easier to handle by finite difference and finite element forward calculations.

The process converges, that is once the smoothest model is found, you can continue to iterate but nothing changes.

bf Most importantly, a problem that had an infinite number of solutions is reduced to one that has a single unique solution: there is only one smoothest model. OK - you could think up ways that would have two equally smooth models that both have the same misfit, but the only time these tend to show up in practice is if the set of acceptable models intersects the line of complete smoothness.

Practical considerations

It is generally best to **start the models from a featureless half-space**, because the Jacobian matrix depends of the conductivity and also directions the direction of the search. If you put in a conductive feature in a MT model where there should not be a conductor, the Jacobian is large within the conductor and it is hard to get rid of it.

For moderately large numbers of data, the expectation value of χ^2 is not very different from the 95% confidence value, both close to RMS=1.

"L-curves" show how misfit and model roughness trade off, but they are not reliable ways of estimating an appropriate misfit, since they depend on how the axes are scaled and how much data are plotted.

The χ^2 quadratic misfit measure is remarkably unforgiving of outliers - the probability of a data point being misfit by 6 error bars is one in a billion. If your data have 2% error bars, a data point that is twice as large as it should be will have the same influence as 2,500 valid data. Check your residuals and remove or down-weight outliers.

Check how your misfit budget is distributed across various parts of your data - RMS misfit is just one number. For example, in MT is the amplitude and phase misfit the same? Is the TE and TM mode misfit the same – in 2D modeling, the TE mode tends to be misfit when there are along-strike variations in resistivity, that is the 2D approximation is breaking down. Does a single site have a much bigger misfit than the rest?

The inversion grid and the computational mesh can be different – this is the "dual grid" approach. Using a dual grid minimizes the size of **J** and reduces both storage and computational time.

The regularization matrix **R** should be scaled with depth and parameter size, or artifacts can develop. Some inversion algorithms, including the popular ModEM 3D MT inversion program, regularize against the starting model, which can have a big effect.

It has been shown (Wheelock et al., 2015) that inverting log(apparent resistivity) is much more stable than linear resistivity or complex impedance. Most 3D codes do not do this because the diagonals of the impedance matrix are zero for the starting half-space model.

References:

Backus G.E., and Gilbert, J.F. , 1967. Numerical applications of a formalism. *Geophysical Journal of the Royal Astronomical Society*, **13**, 247—276.

Constable, S.C., Parker, R.L., and Constable, C.G., 1987. Occam's Inversion: A practical algorithm for generating smooth models from EM sounding data. *Geophysics*, **52**, 289—300.

Marquardt, D.W., 1963. An algorithm for least-squares estimation of non-linear parameters. *Journal of the Society for Industrial and Applied Mathematics*, **11**, 431—441.

Medin, A.E., R.L. Parker, and S. Constable, 2007. Making sound inferences from geomagnetic sounding. *Phys. Earth Planet. Int.*, **160**, 51–59.

Parker, R.L., 1994. *Geophysical Inverse Theory*. Princeton, NJ, Princeton University Press.

Parker, R.L., and K. Whaler, 1981. Numerical methods for establishing solutions to the inverse problem of electromagnetic induction. *Journal of Geophysical Research*, **86**, 9574–9584.

Parker, R.L., and J.R. Booker, 1996. Optimal one-dimensional inversion and bounding of magnetotelluric apparent resistivity and phase measurements. *Physics of the Earth and Planetary Interiors*, **98**, 269–282.

Tikhonov, A.N., and Arsenin, V.Y. , 1977. *Solutions of Ill-Posed Problems*. New York, John Wiley and Sons.

Wheelock, B., S. Constable, and K. Key, 2015. The advantages of logarithmically scaled data for electromagnetic inversion. *Geophysical Journal International*, **201**, 1765–1780.