

CHAPTER 4: LINEARIZATION OF NONLINEAR PROBLEMS

4.1 Introduction

Thus far we have dealt with the linear, explicit forward problem given by

$$\mathbf{G}\mathbf{m} = \mathbf{d} \quad (1.13)$$

where \mathbf{G} is a matrix of coefficients (constants) that multiply the model parameter vector \mathbf{m} and return a data vector \mathbf{d} . If \mathbf{m} is doubled, then \mathbf{d} is also doubled.

We can also write Equation (1.13) out explicitly as

$$d_i = \sum_{j=1}^M G_{ij} m_j \quad i = 1, 2, \dots, N \quad (4.1)$$

This form emphasizes the linear nature of the problem. Next, we consider a more general relationship between data and model parameters.

4.2 Linearization of Nonlinear Problems

Consider a general (explicit) relationship between the i th datum and the model parameters given by

$$d_i = g_i(\mathbf{m}) \quad (4.2)$$

An example might be

$$d_1 = 2m_1^3 \quad (4.3)$$

The steps required to linearize a problem of the form of Equation (4.2) are as follows:

Step 1. Expand $g_i(\mathbf{m})$ about some point \mathbf{m}_0 in model space using a Taylor series expansion:

$$d_i = g_i(\mathbf{m}) \approx g_i(\mathbf{m}_0) + \sum_{j=1}^M \left[\frac{\partial g_i(\mathbf{m})}{\partial m_j} \Big|_{\mathbf{m}=\mathbf{m}_0} \cdot \Delta m_j \right] + \frac{1}{2} \sum_{j=1}^M \left[\frac{\partial^2 g_i(\mathbf{m})}{\partial m_j^2} \Big|_{\mathbf{m}=\mathbf{m}_0} \cdot \Delta m_j^2 \right] + 0(\Delta m_j^3) \quad (4.4)$$

where $\Delta \mathbf{m}$ is the difference between \mathbf{m} and \mathbf{m}_0 , or

$$\Delta \mathbf{m} = \mathbf{m} - \mathbf{m}_0 \quad (4.5)$$

If we assume that terms in Δm_j^n , $n \geq 2$, are small with respect to Δm_j terms, then

$$d_i = g_i(\mathbf{m}) \approx g_i(\mathbf{m}_0) + \sum_{j=1}^M \left[\frac{\partial g_i(\mathbf{m})}{\partial m_j} \Big|_{\mathbf{m}=\mathbf{m}_0} \cdot \Delta m_j \right] \quad (4.6)$$

Step 2. The predicted data \hat{d}_i at $\mathbf{m} = \mathbf{m}_0$ are given by

$$\hat{d}_i = g_i(\mathbf{m}_0) \quad (4.7)$$

Therefore

$$d_i - \hat{d}_i \approx \sum_{j=1}^M \left[\frac{\partial g_i(\mathbf{m})}{\partial m_j} \Big|_{\mathbf{m}=\mathbf{m}_0} \cdot \Delta m_j \right] \quad (4.8)$$

Step 3. We can define the misfit Δc_i as

$$\Delta c_i = d_i - \hat{d}_i \quad (4.9)$$

= observed data – predicted data

Δc_i is *not necessarily* noise. It is just the misfit between observed and predicted data for some choice of the model parameter vector \mathbf{m}_0 .

Step 4. The partial derivative of the i th data equation with respect to the j th model parameter is given by

$$\frac{\partial g_i(\mathbf{m})}{\partial m_j} \quad (4.10)$$

These partial derivatives are *functions* of the model parameters and may be nonlinear (gasp) or occasionally even nonexistent (shudder).

Fortunately, the values of these partial derivatives, evaluated at some point in model space \mathbf{m}_0 , and given by

$$\left. \frac{\partial g_i(\mathbf{m})}{\partial m_j} \right|_{\mathbf{m}=\mathbf{m}_0} \quad (4.11)$$

are just numbers (constants), if they exist, and not functions. We then define G_{ij} as follows:

$$G_{ij} = \left. \frac{\partial g_i(\mathbf{m})}{\partial m_j} \right|_{\mathbf{m}=\mathbf{m}_0} \quad (4.12)$$

Step 5. Finally, combining the above we have

$$\Delta c_i = \sum_{j=1}^M G_{ij} \Delta m_j \Big|_{\mathbf{m}=\mathbf{m}_0} \quad i = 1, \dots, N \quad (4.13)$$

or, in matrix notation, the linearized problem becomes

$$\Delta \mathbf{c} = \mathbf{G} \Delta \mathbf{m} \quad (4.14)$$

where

$$\begin{aligned} \Delta c_i &= d_i - \hat{d}_i = \text{observed data} - \text{predicted data} \\ &= d_i - g_i(\mathbf{m}_0) \end{aligned} \quad (4.15)$$

$$G_{ij} = \left. \frac{\partial g_i(\mathbf{m})}{\partial m_j} \right|_{\mathbf{m}=\mathbf{m}_0} \quad (4.16)$$

and

$$\Delta m_j = \text{change from } (\mathbf{m}_0)_j \quad (4.17)$$

Thus, by linearizing Equation (4.2), we have arrived at a set of linear equations, where now Δc_i (the difference between observed and predicted data) is a linear function of changes in the model parameters from some starting model.

Some general comments on Equation (4.14):

1. In general, Equation (4.14) only holds in the neighborhood of \mathbf{m}_0 , and for small changes $\Delta \mathbf{m}$. The region where the linearization is valid depends on the smoothness of $g_i(\mathbf{m})$.

2. Note that \mathbf{G} now changes with each iteration. That is, one may obtain a different \mathbf{G} for each spot in solution space. Having to reform \mathbf{G} at each step can be very time (computer) intensive, and often one uses the same \mathbf{G} for more than one iteration.

4.3 General Procedure for Nonlinear Problems

Step 1. Pick some starting model vector \mathbf{m}_0 .

Step 2. Calculate the predicted data vector $\hat{\mathbf{d}}$ and form the misfit vector

$$\Delta\mathbf{c} = \mathbf{d} - \hat{\mathbf{d}} \quad (4.18)$$

Step 3. Form

$$G_{ij} = \left. \frac{\partial g_i(\mathbf{m})}{\partial m_j} \right|_{\mathbf{m}=\mathbf{m}_0} \quad (4.19)$$

Step 4. Solve for $\Delta\mathbf{m}$ using any appropriate inverse operator (i.e., least squares, minimum length, weighted least squares, etc.)

Step 5. Form a new model parameter vector

$$\mathbf{m}_1 = \mathbf{m}_0 + \Delta\mathbf{m} \quad (4.20)$$

One repeats Steps 1–5 until $\Delta\mathbf{m}$ becomes sufficiently small (convergence is obtained), or *until* $\Delta\mathbf{c}$ becomes sufficiently small (acceptable misfit), or until a maximum number of iterations (failsafe). Note that \mathbf{m}_i (note the boldfaced \mathbf{m}) is the estimate of the model parameters at the i th iteration, and not the i th component on the model parameter vector.

4.4 Three Examples

4.4.1 A Linear Example

Suppose $g_i(\mathbf{m}) = d_i$ is linear and of the form

$$2m_1 = 4 \quad (4.21)$$

With only one equation, we have $\mathbf{G} = [2]$, $\mathbf{m} = [m_1]$, and $\mathbf{d} = [4]$. (I know, I know. It's easy!)
Then

$$\partial d_1 / \partial m_1 = G_{11} = 2 \quad (\text{for all } m_1) \quad (4.22)$$

Suppose that the initial estimate of the model vector $\mathbf{m}_0 = [0]$. Then $\hat{\mathbf{d}} = \mathbf{G}\mathbf{m}_0 = [2][0] = [0]$ and we have

$$\Delta \mathbf{c} = \mathbf{d} - \hat{\mathbf{d}} = [4] - [0] = [4] \quad (4.23)$$

or the change in the first and only element of our misfit vector $\Delta c_1 = 4$. Looking at our lone equation then,

$$G_{11} \Delta m_1 = \Delta c_1 \quad (4.24)$$

$$\text{or} \quad 2\Delta m_1 = 4 \quad (4.25)$$

$$\text{or} \quad \Delta m_1 = 2 \quad (4.26)$$

Since this is the only element in our model-change vector [in this case, $(\Delta \mathbf{m}_1)_1 = \Delta m_1$], we have $\Delta \mathbf{m}_1 = [2]$, and our next approximation of the model vector, \mathbf{m}_1 , then becomes

$$\mathbf{m}_1 = \mathbf{m}_0 + \Delta \mathbf{m}_1 = [0] + [2] = [2] \quad (4.27)$$

We have just completed Steps 1–5 for the first iteration. Now it is time to update the misfit vector and see if we have reached a solution. Thus, for the predicted data we obtain

$$\hat{\mathbf{d}} = \mathbf{G}\mathbf{m}_1 = [2][2] = [4] \quad (4.28)$$

and for the misfit we have

$$\Delta \mathbf{c} = \mathbf{d} - \hat{\mathbf{d}} = [4] - [4] = [0] \quad (4.29)$$

which indicates that the solution has converged in one iteration. To see that the solution does not depend on the starting point if Equation (4.2) is linear, let's start with

$$(\mathbf{m}_0)_1 = 1000 = m_1 \quad (4.30)$$

Considering the one and only element of our predicted-data and misfit vectors, we have

$$\hat{d}_1 = 2 \times 1000 = 2000 \quad (4.31)$$

$$\text{and} \quad \Delta c_1 = 4 - 2000 = -1996 \quad (4.32)$$

$$\text{then} \quad 2\Delta m_1 = -1996 \quad (4.33)$$

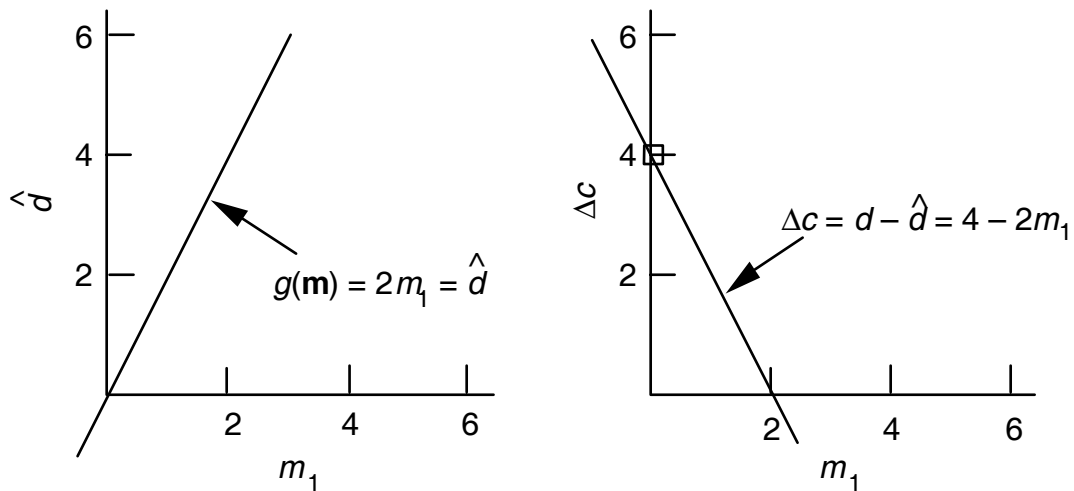
$$\text{or} \quad \Delta m_1 = -998 \quad (4.34)$$

Since Δm_1 is the only element of our first model-change vector, $\Delta \mathbf{m}_1$, we have $\Delta \mathbf{m}_1 = [-998]$, and therefore

$$\mathbf{m}_1 = \mathbf{m}_0 + \Delta \mathbf{m}_1 = [1000] + [-998] = [2] \tag{4.35}$$

As before, the solution has converged in one iteration. This is a general conclusion if the relationship between the data and the model parameters is linear. This problem also illustrates that the nonlinear approach outlined above works when $g_i(\mathbf{m})$ is linear.

Consider the following graphs for the system $2m_1 = 4$:



We note the following:

1. For \hat{d} versus m_1 , we see that the slope $\partial g(m_1)/\partial m_1 = 2$ for all m_1 .
2. For our initial guess of $(\mathbf{m}_0)_1 = 0$, $\Delta c = 4 - 0 = 4$, denoted by a square symbol on the plot of Δc versus m_1 . We extrapolate back down the slope to the point (m_1) where $\Delta c = 0$ to obtain our answer.
3. Because the slope does not change (the problem is linear), the starting guess \mathbf{m}_0 has no effect on the final solution. We can always get to $\Delta c = 0$ in one iteration.

4.4.2 A Nonlinear Example

Now consider, as a second example of the form $g_1(\mathbf{m}) = d_1$, the following:

$$2m^3 = 16 \tag{4.36}$$

Since we have only one unknown, I chose to drop the subscript. Instead, I will use the subscript to denote the iteration number. For example, m_3 will be the estimate of the model parameter m at the third iteration. Note also that, by inspection, $m = 2$ is the solution.

Working through this example as we did the last one, we first note that G_{11} , at the i th iteration, will be given by

$$G_{11} = \left. \frac{\partial g(m)}{\partial m} \right|_{m=m_i} = 3 \times 2m_i^2 = 6m_i^2 \quad (4.37)$$

Note also that G_{11} is now a function of m .

Iteration 1. Let us pick as our starting model

$$m_0 = 1 \quad (4.38)$$

then

$$G_{11} = \left. \frac{\partial g(m)}{\partial m} \right|_{m=m_0} = 6m_0^2 = 6 \quad (4.39)$$

also

$$\hat{d} = 2 \times 1^3 = 2 \quad (4.40)$$

and

$$\Delta c = d - \hat{d} = 16 - 2 = 14$$

Because we have only one element in our model change vector, we have $\Delta \mathbf{c} = [14]$, and the length squared of $\Delta \mathbf{c}$, $\|\Delta \mathbf{c}\|^2$, is given simply by $(\Delta c)^2$:

$$(\Delta c)^2 = 14 \times 14 = 196 \quad (4.41)$$

Now, we find Δm_1 , the change to m_0 , as

$$6\Delta m_1 = 14 \quad (4.42)$$

and

$$\Delta m_1 = 14/6 = 2.3333 \quad (4.43)$$

Thus, our estimate of the model parameter at the first iteration, m_1 , is given by

$$m_1 = m_0 + \Delta m_1 = 1 + 2.3333 = 3.3333 \quad (4.44)$$

Iteration 2. Continuing,

$$G_{11} = 6m_1^2 = 66.66 \quad (4.45)$$

and

$$\hat{d} = 2(3.333)^3 = 74.07 \quad (4.46)$$

thus

$$\Delta c = d - \hat{d} = 16 - 74.07 = -58.07 \quad (4.47)$$

now $(\Delta c)^2 = 3372$ (4.48)

and $66.66\Delta m_2 = -58.07$ (4.49)

gives $\Delta m_2 = -0.871$ (4.50)

thus $m_2 = m_1 + \Delta m_2 = 3.3333 - 0.871 = 2.462$ (4.51)

Iteration 3. Continuing,

$$G_{11} = 6m_2^2 = 36.37 \quad (4.52)$$

and $\hat{d} = 29.847$ (4.53)

thus $\Delta c = -13.847$ (4.54)

now $(\Delta c)^2 = 192$ (4.55)

and $36.37\Delta m_3 = -13.847$ (4.56)

gives $\Delta m_3 = -0.381$ (4.57)

thus $m_3 = m_2 + \Delta m_3 = 2.462 - 0.381 = 2.081$ (4.58)

Iteration 4. (Will this thing ever end??)

$$G_{11} = 6m_3^2 = 25.983 \quad (4.59)$$

and $\hat{d} = 18.024$ (4.60)

thus $\Delta c = -2.024$ (4.61)

now $(\Delta c)^2 = 4.1$ (4.62)

and $25.983\Delta m_4 = -2.024$ (4.63)

gives $\Delta m_4 = -0.078$ (4.64)

thus $m_4 = m_3 + \Delta m_4 = 2.081 - 0.078 = 2.003$ (4.65)

Iteration 5. (When were computers invented???)

$$G_{11} = 6m_4^2 = 24.072 \quad (4.66)$$

and $\hat{d} = 16.072$ (4.67)

thus $\Delta c = -0.072$ (4.68)

now $(\Delta c)^2 = 0.005$ (4.69)

and $24.072\Delta m_5 = -0.072$ (4.70)

gives $\Delta m_5 = -0.003$ (4.71)

thus $m_5 = m_4 + \Delta m_5 = 2.003 - 0.003 = 2.000$ (4.72)

Iteration 6. Beginning, we have

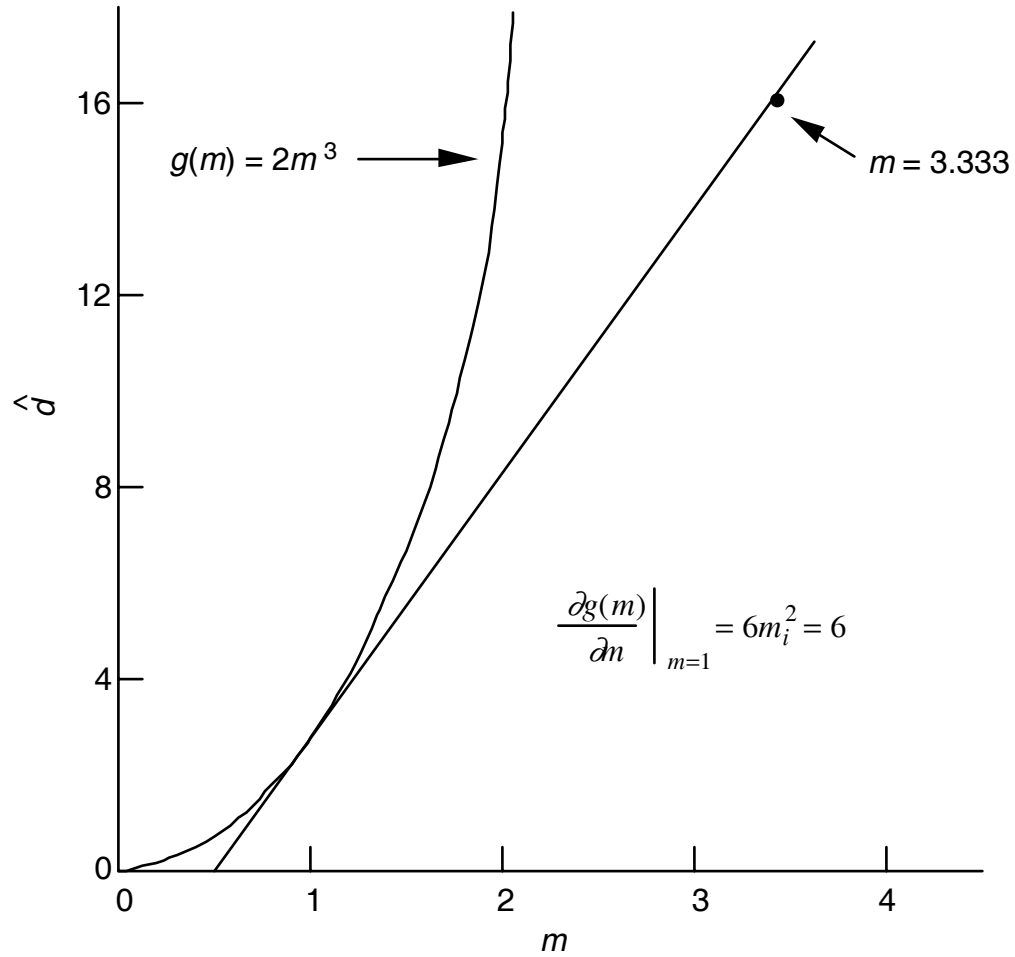
$$G_{11} = 6m_5^2 = 24.000 \quad (4.73)$$

and $\hat{d} = 16.000$ (4.74)

thus $\Delta c = 0.000$ (4.75)

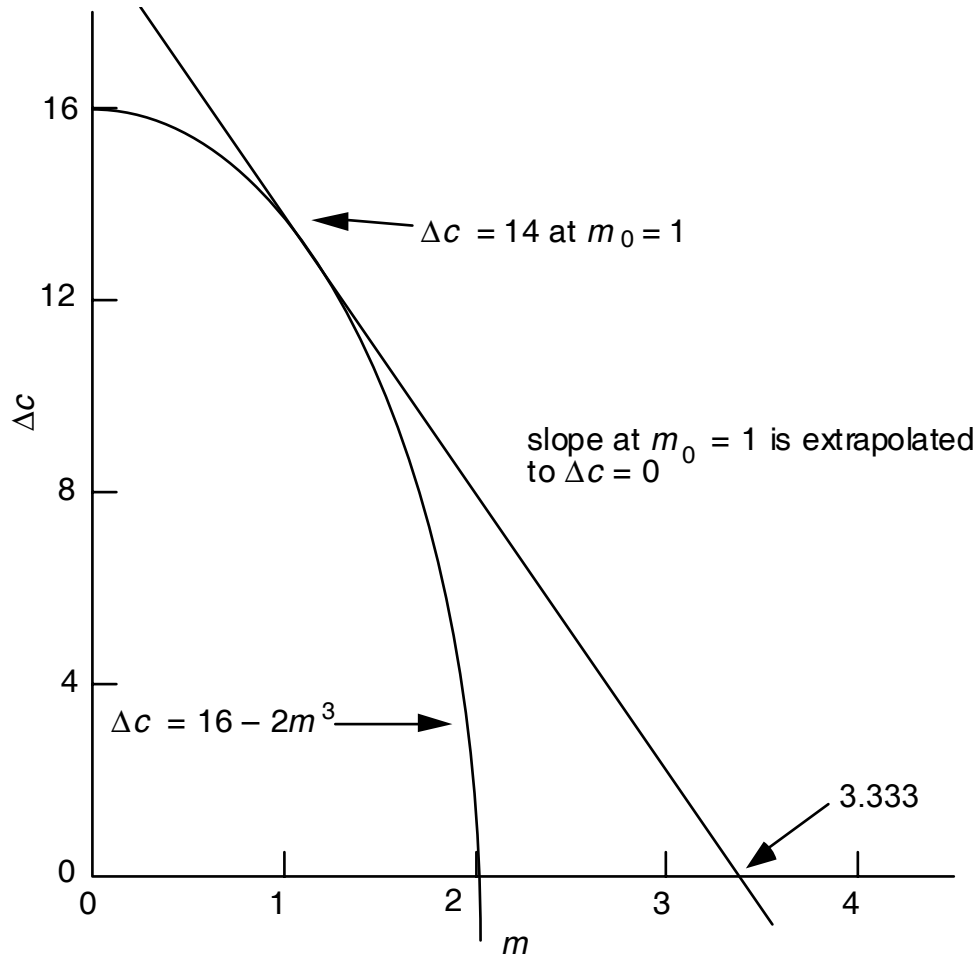
and we quit!!!! We should note that we have quit because the misfit has been reduced to some acceptable level (three significant figures in this case). The solution happens to be an integer, and we have found it to three places. Most solutions are not integers, and we must decide how many significant figures are justified. The answer depends on many things, but one of the most important is the level of noise in the data. If the data are noisy, it does not make sense to claim that a solution to seven places is meaningful.

Consider the following graph for this problem (next page):



Note that the slope at $m = 1$, when extrapolated to give $\hat{d} = 16$, yields $m = 3.333$. The solution then iterates back down the curve to the correct solution $m = 2$.

Consider plotting Δc , rather than \hat{d} , versus m (diagram on next page). This is perhaps more useful because when we solve for Δm_i , we are always extrapolating to $\Delta c = 0$.



For this example, we see that at $m_0 = 1$, the slope $\partial g(m)/\partial m = 6$. We used this slope to extrapolate to the point where $\Delta c = 0$.

At the second iteration, $m_1 = 3.333$ and is farther from the true solution ($m_1 = 2$) than was our starting model. Also, the length squared of the misfit is 3372, much worse than the misfit (196) at our initial guess. This makes an important point: you can still get to the right answer even if some iteration takes you farther from the solution than where you have been. This is especially true for early steps in the iteration when you may not be close to the solution.

Note also that if we had started with m_0 closer to zero, the shallow slope would have sent us to an even higher value for m_1 . We would still have recovered, though (do you see why?). The shallow slope corresponds to a small singular value, illustrating the problems associated with small singular values. We will consider singular-value analysis in the next chapter.

What do you think would happen if you take $m_0 < 0$???? Would it still converge to the correct solution? Try $m_0 = -1$ if your curiosity has been piqued!

4.4.3 Nonlinear Straight-Line Example

An interesting nonlinear problem is fitting a straight line to a set of data points (y_i, z_i) which may contain errors, or noise, along *both* the y and z axes. One could cast the problem as

$$y_i = a + bz_i \quad i = 1, \dots, N \quad (4.76)$$

Assuming z were perfectly known, one obtains a solution for a, b by a linear least squares approach [see Equations (3.32) and (3.37)–(3.39)].

Similarly, if y were perfectly known, one obtains a solution for

$$z_i = c + dy_i \quad i = 1, \dots, N \quad (4.77)$$

again using (3.32) and (3.37)–(3.39). These two lines can be compared by rewriting (4.77) as a function of y , giving

$$y_i = -(c/d) + (1/d)z_i \quad i = 1, \dots, N \quad (4.78)$$

In general, $a \neq -c/d$ and $b \neq 1/d$ because in (4.76) we assumed all of the error, or misfit, was in y , while in (4.77) we assumed that all of the error was in z . Recall that the quantity being minimized in (4.76) is

$$\begin{aligned} E_1 &= [\mathbf{y}^{\text{obs}} - \mathbf{y}^{\text{pre}}]^T [\mathbf{y}^{\text{obs}} - \mathbf{y}^{\text{pre}}] \\ &= \sum_{i=1}^N (y_i - \hat{y}_i)^2 \end{aligned} \quad (4.79)$$

where \hat{y}_i is the predicted y value. The comparable quantity for (4.77) is

$$\begin{aligned} E_2 &= [\mathbf{z}^{\text{obs}} - \mathbf{z}^{\text{pre}}]^T [\mathbf{z}^{\text{obs}} - \mathbf{z}^{\text{pre}}] \\ &= \sum_{i=1}^N (z_i - \hat{z}_i)^2 \end{aligned} \quad (4.80)$$

where \hat{z}_i is the predicted z value.

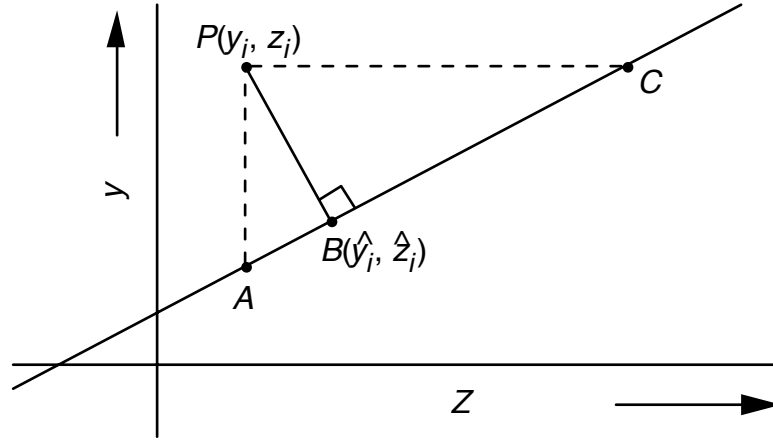
For the best fit line in which both y and z have errors, the function to be minimized is

$$E = \begin{bmatrix} \mathbf{y} - \hat{\mathbf{y}} \\ \mathbf{z} - \hat{\mathbf{z}} \end{bmatrix}^T \begin{bmatrix} \mathbf{y} - \hat{\mathbf{y}} \\ \mathbf{z} - \hat{\mathbf{z}} \end{bmatrix} \quad (4.81)$$

$$E = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2 \quad (4.82)$$

where $y - \hat{y}$ and $z - \hat{z}$ together compose a vector of dimension $2N$ if N is the number of pairs (y_i, z_i) .

Consider the following diagram:



Line PA above represents the misfit in y (see 4.79), line PC represents the misfit in z (see 4.80), while line PB represents the misfit for the combined case (see 4.82).

In order to minimize (4.82) we must be able to write the forward problem for the predicted data (\hat{y}_i, \hat{z}_i) . Let the solution we seek be given by

$$y = m_1 + m_2 z \tag{4.83}$$

Line PB is perpendicular to (4.83), and thus has a slope of $-1/m_2$. The equation of a line through $P(y_i, z_i)$ with slope $-1/m_2$ is given by

$$y - y_i = -(1/m_2)(z - z_i) \tag{4.84}$$

or

$$y = (1/m_2)z_i + y_i - (1/m_2)z \tag{4.85}$$

The point $B(\hat{y}_i, \hat{z}_i)$ is thus the intersection of the lines given by (4.83) and (4.85). Equating the right-hand sides of (4.83) and (4.85) for $y = \hat{y}_i$ and $z = \hat{z}_i$ gives

$$m_1 + m_2 \hat{z}_i = (1/m_2)z_i + y_i - (1/m_2)\hat{z}_i \tag{4.86}$$

which can be solved for \hat{z}_i as

$$\hat{z}_i = \frac{-m_1 m_2 + z_i + m_2 y_i}{1 + m_2^2} \tag{4.87}$$

Rearranging (4.83) and (4.85) to give z as a function of y and again equating for $y = \hat{y}_i$ and $z = \hat{z}_i$ gives

$$(1/m_2)(\hat{y}_i - m_1) = -m_2\hat{y}_i + z_i + m_2y_i \quad (4.88)$$

which can be solved for \hat{y}_i as

$$\hat{y}_i = \frac{m_1 + m_2z_i + m_2^2y_i}{1 + m_2^2} \quad (4.89)$$

substituting (4.89) for \hat{y}_i and (4.87) for \hat{z}_i into (4.82) now gives E as a function of the unknowns m_1 and m_2 . The approach used for the linear problem was to take partials of E with respect to m_1 and m_2 , set them equal to zero, and solve for m_1 and m_2 , as was done in (4.10) and (3.13)–(3.14). Unfortunately, the resulting equations for partials of E with respect to m_1 and m_2 in (4.82) are not linear in m_1 and m_2 and cannot be cast in the linear form

$$\mathbf{G}^T\mathbf{G}\mathbf{m} = \mathbf{G}^T\mathbf{d} \quad (3.21)$$

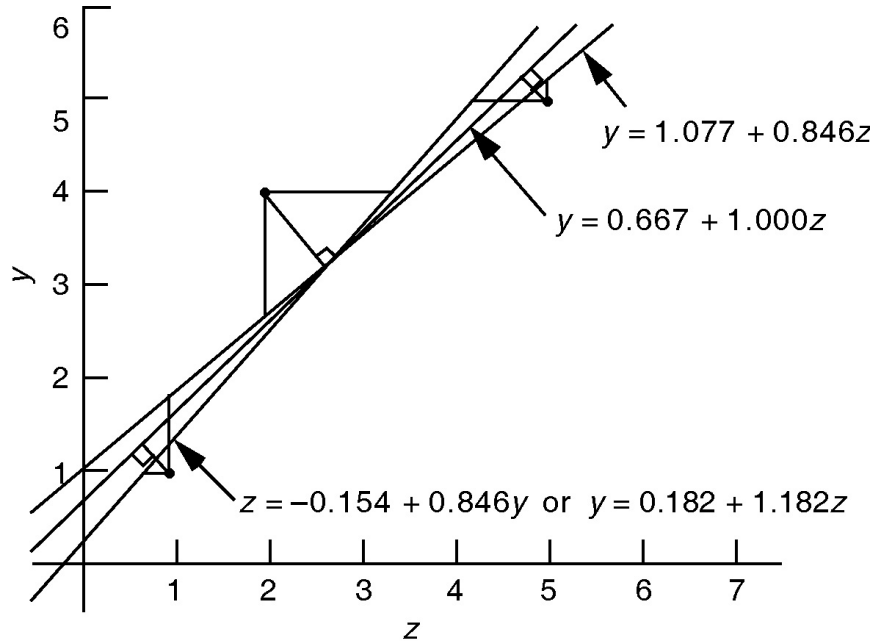
Instead, we must consider (4.87) and (4.89) to be of the form of (4.2):

$$d_i = g_i(\mathbf{m}) \quad (4.2)$$

and linearize the problem by expanding (4.87) and (4.89) in a Taylor series about some starting guesses \hat{z}_0 and \hat{y}_0 , respectively. This requires taking partials of \hat{y}_i and \hat{z}_i with respect to m_1 and m_2 , which can be obtained from (4.87) and (4.89).

Consider the following data set, shown also on the following diagram (next page).

$y_i:$	1	4	5
$z_i:$	1	2	5



The linear least square solution to (4.76)

$$y_i = a + bz_i \quad i = 1, 2, 3 \tag{4.76}$$

is

$$y_i = 1.077 + 0.846z_i \quad i = 1, 2, 3 \tag{4.90}$$

The linear least squares solution to (4.77)

$$z_i = c + dy_i \quad i = 1, 2, 3 \tag{4.77}$$

is

$$z_i = -0.154 + 0.846y_i \quad i = 1, 2, 3 \tag{4.91}$$

For comparison with a and b above, we can rewrite (4.91) with y as a function of z as

$$y_i = 0.182 + 1.182z_i \quad i = 1, 2, 3 \tag{4.92}$$

The nonlinear least squares solution which minimizes

$$E = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2 \tag{4.2}$$

is given by

$$y_i = 0.667 + 1.000z_i \quad i = 1, 2, 3 \quad (4.93)$$

From the figure you can see that the nonlinear solution lies between the other two solutions.

It is also possible to consider a weighted nonlinear least squares best fit to a data set. In this case, we form a new E , after (3.83), as

$$E = \begin{bmatrix} \mathbf{y} - \hat{\mathbf{y}} \\ \mathbf{z} - \hat{\mathbf{z}} \end{bmatrix}^T \mathbf{W}_e \begin{bmatrix} \mathbf{y} - \hat{\mathbf{y}} \\ \mathbf{z} - \hat{\mathbf{z}} \end{bmatrix} \quad (4.94)$$

and where \mathbf{W}_e is a $2N \times 2N$ weighting matrix. The natural choice for \mathbf{W}_e is

$$\left\{ \text{cov} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \right\}^{-1}$$

the inverse data covariance matrix. If the errors in y_i , z_i are uncorrelated, then the data covariance matrix will be a diagonal matrix with the variances for y_i as the first N entries and the variances for z_i as the last N entries. If we further let V_y and V_z be the variances for y_i and z_i , respectively, then Equations (4.87) and (4.89) become

$$\hat{z}_i = \frac{-m_1 m_2 V_z + V_y z_i + m_2 V_z y_i}{m_2^2 V_z + V_y} \quad (4.95)$$

and

$$\hat{y}_i = \frac{m_1 V_y + m_2 V_y z_i + m_2^2 V_z y_i}{m_2^2 V_z + V_y} \quad (4.96)$$

If $V_z = V_y$, then dividing through either (4.95) or (4.96) by the variance returns (4.87) or (4.89). Thus we see that weighted least squares techniques are equivalent to general least squares techniques when all the data variances are equal and the errors are uncorrelated, as expected.

Furthermore, dividing both the numerator and denominator of (4.96) by V_y yields

$$\hat{y}_i = \frac{m_1 + m_2 z_i + (m_2^2 V_z y_i) / V_y}{[(m_2^2 V_z) / V_y] + 1} \quad (4.97)$$

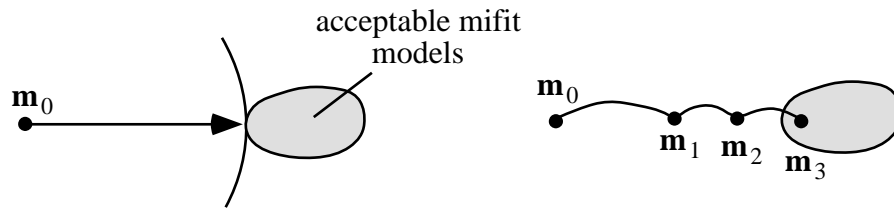
Then, in the limit that V_z goes to zero, (4.97) becomes

$$\hat{y}_i = m_1 + m_2 z_i \quad (4.98)$$

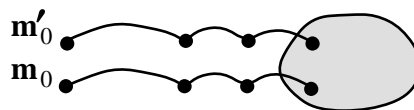
which is just the linear least squares solution of (4.76). That is, if z_i is assumed to be perfectly known ($V_z = 0$), the nonlinear problem reduces to a linear problem. Similar arguments can be made for (4.95) to show that the linear least squares solution of (4.77) results when V_y goes to zero.

4.5 Creeping vs. Jumping (Shaw and Orcutt, 1985)

The general procedure described in Section 4.3 is termed "creeping" by Shaw and Orcutt (Shaw, P. R., and Orcutt, J. A., Waveform inversion of seismic refraction data and applications to young Pacific crust, *Geophys. J. Roy. Astr. Soc.*, 82, 375-414, 1985). It finds, from the set of acceptable misfit models, the one closest to the starting model under the Euclidian norm.



Because of the nonlinearity, often several iterations are required to reach the desired misfit. The acceptable level of misfit is free to vary and can be viewed as a parameter that controls the trade-off between satisfying the data and keeping the model perturbation small. Because the starting model itself is physically reasonable, the unphysical model estimates tend to be avoided. There are several potential disadvantages to the creeping strategy. Creeping analysis depends significantly on the choice of the initial model. If the starting model is changed slightly, a new final model may well be found. In addition, constraints applied to model perturbations may not be as meaningful as those applied directly to the model parameters.



Parker (1994) introduced an alternative approach with a simple algebraic substitution (Parker, R. L., *Geophysical Inverse Theory*, Princeton University Press, 1994). The new method, called "jumping," directly calculates the new model in a single step rather than calculating a perturbation to the initial model. Now, any suitable norm can be applied to the model rather than to the perturbations.

This new strategy is motivated, in part, by the desire to map the neighborhood of starting models near \mathbf{m}_0 to a single final model, thus making the solution less sensitive to small change in \mathbf{m}_0 .



Let's write the original nonlinear equations as

$$\mathbf{gm} = \mathbf{d} \quad (4.99)$$

After linearization about an initial model \mathbf{m}_0 , we have

$$\mathbf{G}\Delta\mathbf{m} = \Delta\mathbf{c} \quad (4.100)$$

when

$$\mathbf{G} = \left. \frac{\partial \mathbf{g}}{\partial \mathbf{m}} \right|_{\mathbf{m}_0}, \quad \Delta\mathbf{c} = \mathbf{d} - \mathbf{gm}_0 \quad (4.101)$$

The algebraic substitution suggested by Parker is to simply add \mathbf{Gm}_0 to both sides, yielding

$$\mathbf{G}\Delta\mathbf{m} + \mathbf{Gm}_0 = \Delta\mathbf{c} + \mathbf{Gm}_0$$

then

$$\mathbf{G}[\Delta\mathbf{m} + \mathbf{m}_0] = \Delta\mathbf{c} + \mathbf{Gm}_0 \quad (4.102)$$

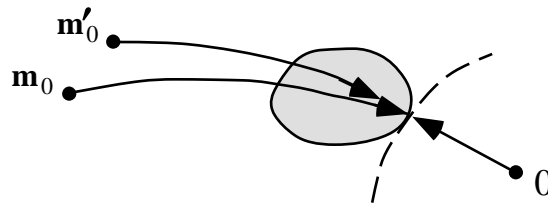
and

$$\boxed{\mathbf{Gm}_1 = \Delta\mathbf{c} + \mathbf{Gm}_0} \quad (4.103)$$

or

$$\mathbf{Gm}_1 = \mathbf{d} - \mathbf{gm}_0 + \mathbf{Gm}_0 \quad (4.104)$$

At this point, this equation is algebraically equivalent to our starting linearized equation. But the crucial difference is that now we are solving directly for the model \mathbf{m} rather than a perturbation $\Delta\mathbf{m}$. This slight algebraic difference means we can now apply any suitable constraint to the model. A good example is a smoothing constraint. If the initial model is not smooth, applying a smoothing constraint to the model perturbations may not make sense. In the new formulation, we can apply the constraint directly to the model. In the jumping scheme, the new model is computed directly, and the norm of this model is minimized relative to an absolute origin 0 corresponding to this norm.

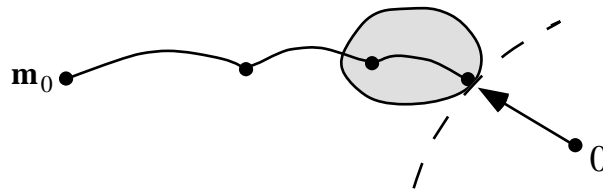


The explicit dependence on the starting model is greatly reduced.

In our example of the second derivative smoothing matrix \mathbf{D} , we can now apply this directly to the jumping equations.

$$\begin{bmatrix} \mathbf{G} \\ \theta \mathbf{D} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \Delta \mathbf{c} + \mathbf{G} \mathbf{m}_0 \\ \mathbf{0} \end{bmatrix} \quad (4.105)$$

We should keep in mind that since the problem is nonlinear, there is still no guarantee that the final model will be unique, and even this "jumping" scheme is iterative.



To summarize, the main advantage of the jumping scheme is that the new model is calculated directly. Thus, constraints can be imposed directly on the new model. The minimization of the squared misfit can be traded off with the constraint measures, allowing optimization of some physically significant quantity. This tends to reduce the strong dependence on the initial model that is associated with the creeping scheme.

We now turn our attention to the generalized inverse in the next three chapters. We will begin with eigenvalue problems, and then continue with singular-value problems, the generalized inverse, and ways of quantifying the quality of the solution.