STEP SIZE IN OPTIMIZATION PROBLEMS

Jon F. Claerbout

In non-linear iterative optimization we jump along in the direction of the gradient. Estimates of the distance to the gradient must be based on either (1) a good understanding of the physics involved, or (2) prior estimates of the second derivatives, which are known as the covariance, the Hessian, or the Levinson, matrix. The distance estimates are troublesome for reasons that will be given; a suggestion for managing them will be made.

Spatial Non-constancy of Second Derivatives

The usual least-squares problem is to vary x in order to minimize

$$E = (d - ax)^2 \tag{1}$$

Note that

$$\frac{\partial^2 E}{\partial x^2} = a^2 = \text{constant function of } x \tag{2}$$

The final results of many practical problems are not significantly different if we minimize absolute errors:

$$E = |d - ax| \tag{3}$$

Note further that

$$\frac{\partial E}{\partial x} = -a \operatorname{sgn}(d - ax) \tag{4}$$

$$\frac{\partial^2 E}{\partial x^2} = a^2 \delta(d - ax) \neq \text{constant function of } x$$

It is important that an apparently minor change in problem formulation converts a constant function to a delta function.

More general non-linear optimization problems are usually approached by a Taylor series that is truncated at second order and ends up looking like equation (1).

Stochastic Behavior of Second Derivatives

In solving a filtering problem in which there are 10^6 observations, it seems quite reasonable to find 100 filter coefficients, since the gradient of the error norm is a 100-component vector. The second derivative matrix (covariance, Levinson, or Hessian matrix), however, has 100 x 100 or 10^4 elements. Does this make sense, in view of the fact that the square root of the number of data points is 10^3?

The appropriate philosophy could be to somehow reduce the number of degrees of freedom in the second derivative. We can reduce the dimensionality to one if we introduce a new variable α that scales the gradient vector $g = \partial E/\partial x$. Now we can forget about the second derivative of matrix E with respect to the vector x, and think only of the second derivative scalar of E with respect to α.

Following a less extreme philosophy, we could include α times the gradient, plus β times the previous gradient, and then use the distance implied by this two-by-two covariance matrix.

A Suggested Method

Newton's method, drawn from a Taylor series to second order with the gradient zeroed, is

$$dx = \frac{-\frac{\partial E}{\partial x}}{\frac{\partial^2 E}{\partial x^2}} \tag{6}$$

In the multivariate problem the second derivative is a matrix. A better approach than using the second derivative would be to use the second finite difference operator. It averages the second derivative over a region of interest, reducing the space-variability problem. Think about descending through the continental drainage pattern: Over what dx should we do the differencing? Obviously, over the dx that separates the last two function evaluations. Let g_t denote the gradient determined at the present place and g_{t-1} denote the previous gradients. For a scalar problem we replace equation (6) by

$$dx_{t+\frac{1}{2}} = \frac{-g_t}{\left[\frac{g_t - g_{t-1}}{dx_{t-\frac{1}{2}}}\right]} \tag{7a}$$

Let us abbreviate this expression:

$$dx' = \frac{g_t}{\left[\frac{g_t - g_{t-1}}{dx}\right]} \tag{7b}$$

For the multivariate case, a good suggestion is

$$dx' = \frac{-g_t \, (dx \cdot dx)}{(g_t \cdot dx) - (g_{t-1} \cdot dx)} \tag{8}$$

We may interpret (8) as follows:

(1) If g_t is approximately equal to the negative of g_{t-1}, the step size will be halved. This is good.

(2) If g_t is approximately equal to g_{t-1}, the step size will be drastically increased. The results will probably be good.

(3) If the denominator is negative we are on a non-convex surface. This step size estimate is worthless.

In the Case of a Non-convex Surface

Some physics offers the best help here. Otherwise, decide how many iterations you can afford, and decrease |dx|/|x| linearly at each step.