### Notes for 2016-04-29

# Parameter Fitting

Our focus this week is on optimization problems and nonlinear equations with specialized structure. Last lecture, we touched on some of the many places where spectral methods make sense. This time, we consider nonlinear least squares problems. Such problems are common in cases where we want to fit model parameters to (possibly inconsistent) data, and there are a variety of methods to treat them, taking advantage of different types of structure in the problem.

# Gauss-Newton

Suppose $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $m > n$, and let

$$\phi(x) = \frac{1}{2}\|F(x)\|^2.$$

We consider the nonlinear least squares problem of minimizing $\phi$. The most obvious approach, supposing we are willing to compute lots of derivatives, is to apply Newton iteration. To do this, we need not only function values, but also gradients and Hessians:

$$\begin{aligned}
\phi(x) &= \frac{1}{2}\sum_{i=1}^m F_i(x)^2 \\
\phi'(x) &= \sum_{i=1}^m F_i(x)F_i'(x) \\
\phi''(x) &= \sum_{i=1}^m F_i'(x)^T F_i'(x) + F_i(x)F_i''(x)
\end{aligned}$$

In terms of the residual $F(x) \in \mathbb{R}^m$ and the Jacobian $J(x) = F'(x) \in \mathbb{R}^{m \times n}$,

$$\begin{aligned}
\nabla\phi(x) &= J^T F \\
\phi''(x) = H &= J^T J + \sum_{i=1}^m F_i(x)F_i''(x).
\end{aligned}$$

That second Hessian term is awkward, as it involves computing a Hessian of each component of $F$ in terms – that's a lot of derivatives! Fortunately, if the solution has small residual (each $F_i$ is small at the optimum), then the second term in the Hessian also should not matter very much. If we drop the second term, we get the *Gauss-Newton iteration*

$$x^{k+1} = x^k - (J^T J)^{-1} J^T F(x^k)$$

which, having absorbed the first half of the semester, we identify as

$$x^{k+1} = x^k - J^\dagger F(x^k)$$

where $J^\dagger$ is the Moore-Penrose pseudo-inverse (a.k.a. the solution operator for a linear least squares problem). Equivalently, we say that $x^{k+1} = x^k + p^k$ where $p^k$ is the solution to

$$\text{minimize } \|F(x^k) + J(x^k)p^k\|^2.$$

That is, the Gauss-Newton iteration can be seen either as dropping an inconvenient term in a Newton iteration, or as successive miminizing linear least-squares models of the nonlinear least-squares problem.

## Convergence of Gauss-Newton and Beyond

Unlike Newton, Gauss-Newton has the attractive feature that (assuming $J$ is full rank) it always produces a descent direction at any point that is not already a stationary point:

$$-\nabla\phi(x)^T p = -(J^T F)^T (J^\dagger F) = -F^T \Pi F.$$

where $\Pi = JJ^\dagger$ is the orthogonal projector onto the range space of $J$.

What of the asymptotic convergence behavior? With some work, we can get an error iteration

$$\|e^{\text{new}}\| \lesssim \frac{M\|F(x_*)\|}{\sigma_{\min}(J(x_*))^2} \|e^{\text{old}}\|,$$

where $M$ is a local Lipschitz constant for $J$. This shows local linear convergence if $M\|F(x_*)\| < \sigma_{\min}(J(x_*))^2$. But if $F(x_*)$ is large relative to the small singular values of $J(x_*)$, the iteration *will not converge without safeguards*.

We can restore guaranteed convergence by a line search. Alternately, we can use a trust region approach or regularization term in the linear least squares problem for the update; this leads to the Levenberg-Marquardt iteration.

## Iteratively Reweighted Least Squares

Many nonlinear least squares problems involve a nonlinear loss function applied elementwise to a linear residual. We assume the loss is positive, and write it as $\ell(r_i) = f(r_i)^2$, so the problem becomes

$$\text{minimize } \frac{1}{2} \sum_i \ell(r_i) = \frac{1}{2} \sum_i f(r_i)^2, \quad r = Ax - b.$$

This is a nonlinear least squares problem; if we apply the Gauss-Newton idea, we have steps of the form

$$\text{minimize } \|f(r) + \operatorname{diag}(f'(r))Ap\|^2.$$

That is, each step is the solution to a weighted least squares problem

$$\text{minimize } \|W(Ap - b)\|^2$$

where the diagonal weight matrix $W$ has $w_{ii} = f'(r_i)$ and $b_i = -f(r_i)/f'(r_i)$. Because the weights vary at each step, this is known as an *iteratively reweighted least squares* (IRLS) method.

The iteratively reweighted least squares idea also appears in statistics under the guise of *Fisher scoring* for computing maximum likelihood estimates. As with Gauss-Newton, the idea behind Fisher scoring is to replace a hard-to-manage Hessian with something simpler (the distributional expected value of the Hessian as opposed to the Hessian derived from the sample). Unfortunately, the literature is often slightly confusing in that many authors fail to distinguish an exact Newton iteration on the scoring function from the approximate Newton iteration in Fisher's approach.

## Variable Projection

The case of a linear model with a non-quadratic loss function leads to one special class of nonlinear least squares. Another common special case is when a model depends linearly on some parameters and nonlinearly on others. For these problems, *variable projection* is the general strategy of eliminating the variables on which a least squares problem depends linearly.

As an example, consider the problem

$$\text{minimize } \phi(x, y) = \frac{1}{2}\|A(y)x - b\|^2$$

where $A : \mathbb{R}^p \rightarrow \mathbb{R}^{m \times n}$ depends (possibly nonlinearly) on $y$, but the $x$ variables only enter the problem linearly. The *variable projection* approach involves eliminating the $x$ variables from the equation:

$$\text{minimize } \phi(x(y), y) = \frac{1}{2}\|r(y)\|^2, \quad r(y) = A(y)x(y) - b, \quad x(y) = A(y)^\dagger b.$$

The variation of $\|r\|^2/2$ is $r^T \delta r$ where

$$\delta r = A(\delta x) + (\delta A)x;$$

and because $A^T r = 0$ (normal equations), we have

$$r^T \delta r = r^T (\delta A)x.$$

One may undertake second derivatives as an exercise in algebraic fortitude; alternately, we may apply BFGS or similar methods. Either way, we are now left with a smaller optimization problem involving the $y$ variables alone.