Notes for 2017-05-01

1 Introduction

We briefly discussed nonlinear least squares problems in a previous lecture, when we described the historical path leading to trust region methods starting from the Gauss-Newton and Levenberg-Marquardt iterations. In this lecture, after a recap of the basic setup and of the Gauss-Newton and Levenberg-Marquardt iterations, we discuss more general regression problems and the common *iteratively reweighted least squares* (IRLS) family of algorithms for solving them.

2 Nonlinear least squares algorithms

In a nonlinear least squares problem, we suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ where typically $m > n$ and we seek to solve the problem

$$\text{minimize } \phi(x), \quad \phi(x) \equiv \frac{1}{2}\|f(x)\|^2.$$

Computing derivatives using the chain rule, we have

$$\begin{aligned}
\frac{\partial \phi}{\partial x_i} &= \sum_k f_k \frac{\partial f_k}{\partial x_i} \\
\frac{\partial^2 \phi}{\partial x_i \partial x_j} &= \sum_k \frac{\partial f_k}{\partial x_i} \frac{\partial f_k \partial x_j}{+} f_k \frac{\partial^2 f_k}{\partial x_i \partial x_j},
\end{aligned}$$

or, more tersely (letting $J = f' \in \mathbb{R}^{m \times n}$ denote the Jacobian),

$$\begin{aligned}
\nabla \phi &= J^T f \\
H_\phi &= J^T J + \sum_k f_k H_{f_k}.
\end{aligned}$$

Hence, a Newton step for this problem looks like

$$p_{\text{Newton}} = -\left(J^T J + \sum_k f_k H_{f_k} \right)^{-1} J^T f.$$

Taking a Newton step involves computing Hessians for each component function f_k, which we often would prefer to avoid. If the residual components $f_k(x)$ are all close to zero at the minimizer for ϕ, then the Hessian is $H_\phi \approx J^T J$ near the minimizer; this gives us a *Gauss-Newton* step

$$p_{GN} = -(J^T J)^{-1} J^T f = -J^\dagger f.$$

The Gauss-Newton step can be derived from plugging a Taylor expansion of f in the nonlinear least squares objective:

$$\text{minimize } \frac{1}{2}\|f + Jp\|^2.$$

Indeed, this is the way we derived Gauss-Newton the first time we discussed it. However, as we saw even in our discussion of linear least squares problems, least squares fits may be problematic when the matrix involved is ill-conditioned. The *Levenberg-Marquardt* iteration adds a Tikhonov regularization term to the linear least squares problem for the step:

$$\text{minimize } \frac{1}{2}\|f + Jp\| + \frac{\lambda}{2} p^T D p$$

where D is typically chosen to be I (Levenberg) or $\operatorname{diag}(J^T J)$ (Marquardt). The latter choice is often preferred because it is invariant under re-scaling of the variables in the problem.

Gauss-Newton iteration is not guaranteed to converge at all, though the Gauss-Newton direction is a descent direction, and so the iteration converges to a stationary point (under reasonable hypotheses) with the aid of a line search. The Levenberg-Marquardt iteration similarly converges to a stationary point (under the usual hypotheses) if λ is chosen reasonably; the step behaves like

$$p_{LM} = -\lambda^{-1} \nabla \phi + O(\lambda^{-2})$$

for large λ, and hence a heavily-damped Levenberg-Marquardt step is essentially a simple gradient descent step. For small λ, the Levenberg-Marquardt step converges to the Gauss-Newton step. One can choose λ adaptively by thinking of it as a parameter for a trust-region strategy, or follow a simpler choice of adaptively increasing λ (e.g. $\lambda := \lambda\nu$ for some $\nu > 1$) when a proposed step fails to decrease the objective, or decreasing λ (e.g. $\lambda := \lambda/nu$) after a few (or just one) successful decrease steps.

3 A statistical digression

Nonlinear least squares problems frequently arise in practice because we want to fit parameters of some model based on noisy data. That is, we believe that we have m observed values $y_1, \ldots y_m$ drawn from some distribution, and we would like to recover the parameters of the distribution. The most common variant is

$$Y_i \sim N(\mu_i(\boldsymbol{\beta}), \sigma^2);$$

that is, each observation is assumed to be independent of the others, with means μ_i given according to a parametric model with parameters $\boldsymbol{\beta}$. In this case, the *log likelihood* function for the parameter vector $\boldsymbol{\beta}$ given a data vector y is

$$L(\boldsymbol{\beta}|y) = -\sum_{k=1}^{m} \frac{1}{2\sigma^2}(y_k - \mu_k(\boldsymbol{\beta}))^2 + C,$$

and maximizing the (log) likelihood is equivalent to minimizing the norm of the residual function $f(\boldsymbol{\beta}) = \boldsymbol{\mu}(\boldsymbol{\beta}) - y$.

The statistical interpretation of least squares is relevant not only to understanding the problem, but also as a way of understanding the Gauss-Newton solver algorithm. The log likelihood is a random variable that depends on the random variables $Y_1, \ldots, Y_m$; if we assume each Y_i is drawn from a distribution with mean $\mu_i(\boldsymbol{\beta})$ for a given parameter vector $\boldsymbol{\beta}$, then the *expected* Hessian values with respect to the β variables are

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 L}{\partial\beta_i \partial\beta_j}\right] &= \mathbb{E}\left[\sum_{k,l} \frac{\partial L}{\partial\mu_k}\frac{\partial\mu_k}{\partial\beta_i}\frac{\partial L}{\partial\mu_l}\frac{\partial\mu_l}{\partial\beta_j} + \sum_k \frac{\partial L}{\partial\mu_k}\frac{\partial^2\mu_k}{\partial\beta_i\partial\beta_j}\right] \\
&= \mathbb{E}\left[\sum_{k,l} \frac{\partial L}{\partial\mu_k}\frac{\partial\mu_k}{\partial\beta_i}\frac{\partial L}{\partial\mu_l}\frac{\partial\mu_l}{\partial\beta_j} + \right]
\end{aligned}$$

where the second term vanishes because

$$\frac{\partial L}{\partial\mu_k} = \sigma^{-2}(Y_k - \mu_k(\boldsymbol{\beta}))$$

has expected value equal to zero. If we write the Jacobian of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\beta}$ as the matrix A, the *expected* Hessian is

$$\mathcal{I}(\boldsymbol{\beta}) = A^T\Sigma^{-2}A,$$

where $\Sigma^{-2} = \sigma^{-2} I$ in this case. The expected Hessian is known in statistics as Fisher's information matrix, and using the expected Hessian (or information matrix) in place of the actual Hessian is known as *scoring*. Hence, Gauss-Newton sometimes goes under the names like "Fisher's scoring method."

4 Beyond Gaussian

Maximum likelihood estimation leads to least squares problems when the observations are independent normal random variables; whether the least squares problems are linear or nonlinear depends on whether the mean vector $\boldsymbol{\mu}$ depends linearly or nonlinearly on the parameters $\boldsymbol{\beta}$. But not every random variable is Gaussian! In a more general setting, we might have that the log likelihood takes the form

$$L(\boldsymbol{\beta}|y) = -\sum_{k=1}^{m} \rho(y_k - \mu_k(\boldsymbol{\beta})) + C.$$

where we separate out normalization constants in order to ensure that $\rho(0) = 0$ and $\rho(r) > 0$ for $r \neq 0$. Hence, maximum likelihood estimation has the form

$$\text{minimize } \phi(\boldsymbol{\beta}) = \sum_{k=1}^{m} \rho(f_k(\boldsymbol{\beta}))$$

where $f_k(\boldsymbol{\beta}) = y_k - \mu_k(\boldsymbol{\beta})$ is a residual, and ρ is a loss function[1]. In some cases it is convenient to instead write the likelihood function in terms of a transformed variable, e.g.

$$L(\boldsymbol{\beta}|y) = -\sum_{k=1}^{m} \rho(g(y_k) - \eta_k(\boldsymbol{\beta})) + C.$$

This may be familiar to those of you who have studied logistic regression, for example.

Let us consider the case where $\boldsymbol{\mu} = A\boldsymbol{\beta}$ and the loss function is even (i.e. $\rho(-r) = \rho(r)$) and continuously differentiable. Define ψ to be the deriva-

[1] We could have a different loss function for each data point; but while this would not significantly change the development of the ideas, it would mean tracking one more subscript that I could get wrong. So let's stick to a single loss function, shall we?

tive of ρ; then

$$\begin{aligned}
f &= y - A\boldsymbol{\beta} \\
\phi(\boldsymbol{\beta}) &= \sum_{k=1}^{m} \rho(f_k) \\
\nabla\phi &= A^T \psi(f) \\
H_\phi &= A^T \operatorname{diag}(\psi'(f)) A
\end{aligned}$$

where $\psi(f)$ and $\psi'(f)$ should be interpreted as applying ψ and ψ' elementwise to the vector f (we do not need to wave our hands to get rid of Hessians of components of f in this case, since f is linear in $\boldsymbol{\beta}$). A Newton step for this problem takes the form

$$p = -(A^T \operatorname{diag}(\psi'(f))A)^{-1} A^T \psi(f)$$

which is equivalent to the *weighted* least squares problem

$$\text{minimize } \left\| \frac{\psi(f)}{\psi'(f)} + Ap \right\|^2_{\operatorname{diag}(\psi'(f))}.$$

where $\psi(f)/\psi'(f)$ should be interpreted elementwise.

Of course, we already know that Newton iteration is not globally convergent in general, but this iteration has another irritating issue: it is undefined when any component of $\psi'(f)$ is exactly zero! There are interesting loss functions (such as Huber's loss function, often used in robust regression) where this is a real issue. However, an a slight modification avoids this problem and *is* globally convergent for convex loss functions. Define $W(f)$ to be the diagonal matrix of weights $w_k = \psi(f_k)/f_k$; then $\boldsymbol{\beta}$ is a stationary point for ϕ if

$$A^T \psi(f) = A^T W(f) f = 0.$$

This suggests the fixed point iteration

$$A^T W(f^k) f^{k+1} = 0,$$

i.e.

$$\boldsymbol{\beta}^{k+1} = \operatorname{argmin} \|A\boldsymbol{\beta}^k - y\|^2_{W(f^k)}.$$

That is, at each step we compute a new weighted least squares fit to the data. Observe that

$$w_k = \frac{\psi(f_k)}{f_k} = \frac{\psi(f_k) - \psi(0)}{f_k - 0} = \psi'(f_k) + o(f_k)$$

and

$$\frac{\psi(f_k)}{\psi'(f_k)} = \frac{\psi'(0)f_k}{\psi'(0)} + o(f_k) = f_k + o(f_k);$$

hence, this iteration is similar to Newton iteration when the residuals are small.

This algorithm is an example of an *iteratively reweighted least squares* (IRLS) algorithm. Several algorithms share the IRLS name; all have the property that each iterate is the solution to a weighted least squares problem, where the weights vary from iteration to iteration.