Notes for 2017-04-19

1 Gauss-Newton

Before beginning our (brief) discussion of trust region methods, we first turn to another popular iterative solver: the Gauss-Newton method for nonlinear least squares problems. Given $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for $m > n$, we seek to minimize the objective function

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

The Gauss-Newton approach to this optimization is to approximate f by a first order Taylor expansion in order to obtain a proposed step:

$$p_k = \operatorname{argmin}_p \frac{1}{2}\|f(x_k) + f'(x_k)p\|^2 = -f'(x_k)^\dagger f(x_k).$$

Writing out the pseudo-inverse more explicitly, we have

$$\begin{aligned} p_k &= -[f'(x_k)^T f'(x_k)]^{-1} f'(x_k)^T f(x_k) \\ &= -[f'(x_k)^T f'(x_k)]^{-1} \nabla\phi(x_k). \end{aligned}$$

The matrix $f'(x_k)^T f'(x_k)$ is positive definite if $f'(x_k)$ is full rank; hence, the direction p_k is always a descent direction provided x_k is not a stationary point and $f'(x_k)$ is full rank. However, the Gauss-Newton step is *not* the same as the Newton step, since the Hessian of ϕ is

$$H_\phi(x) = f'(x)^T f'(x) + \sum_{j=1}^m f_j(x) H_{f_j}(x).$$

Thus, the Gauss-Newton iteration can be seen as a modified Newton in which we drop the inconvenient terms associated with second derivatives of the residual functions f_j.

Assuming f' is Lipschitz with constant L, an error analysis about a minimizer x_* yields

$$\|e_{k+1}\| \leq L\|f'(x_*)^\dagger\|^2 \|f(x_*)\| \|e_k\| + O(\|e_k\|^2).$$

Thus, if the optimal residual norm $\|f(x_*)\|$ is small, then from good initial guesses, Gauss-Newton converges nearly quadratically (though the linear term will eventually dominate). On the other had, if $\|f(x_*)\|$ is larger than $\|f'(x_*)^\dagger\|$, then the iteration may not even be locally convergent unless we apply some type of globalization strategy.

2 Regularization and Levenberg-Marquardt

While we can certainly apply line search methods to globalize Gauss-Newton iteration, an alternate proposal due to Levenberg and Marquardt is solve a *regularized* least squares problem to compute the step; that is,

$$p_k = \operatorname{argmin}_p \frac{1}{2}\|f(x_k) + f'(x_k)p\|^2 + \frac{\lambda}{2}\|Dp\|^2.$$

The scaling matrix D may be an identity matrix (per Levenberg), or we may choose $D^2 = \operatorname{diag}(f'(x_k)^T f'(x_k))$ (as suggested by Marquardt).

For $\lambda = 0$, the Levenberg-Marquardt step is the same as a Gauss-Newton step. As λ becomes large, though, we have the (scaled) gradient step

$$p_k = -\frac{1}{\lambda} D^{-2} f(x_k) + O(\lambda^{-2}).$$

Unlike Gauss-Newton with line search, changing the parameter λ affects not only the distance we move, but also the direction.

In order to get both ensure global convergence (under sufficient hypotheses on f, as usual) and to ensure that convergence is not too slow, a variety of methods have been proposed that adjust λ dynamically. To judge whether λ has been chosen too aggressively or conservatively, we monitor the *gain ratio*, or the ratio of actual reduction in the objective to the reduction predicted by the (Gauss-Newton) model:

$$\rho = \frac{\|f(x_k)\|^2 - \|f(x_k + p_k)\|^2}{\|f(x_k)\|^2 - \|f(x_k) + f'(x_k)p_k\|^2}.$$

If the step decreases the function value enough (ρ is sufficiently positive), then we accept the step; otherwise, we reject it. For the next step (or the next attempt), we may increase or decrease the damping parameter λ depending on whether ρ is close to one or far from one.

3 Consider constraints

There is another way to think of the Levenberg-Marquardt step. Consider the minimization problem

$$p_k = \operatorname{argmin}_p \frac{1}{2}\|f(x) + f'(x)p\|^2 \text{ s.t. } \|Dp\| \leq \Delta.$$

There are two possible cases in this problem:

1. $\|f'(x_k)^\dagger f(x)\| < \Delta$, and the solution is the Gauss-Newton step
2. Otherwise the Gauss-Newton step is too big, and we have to enforce the constraint $\|Dp\| = \Delta$. For convenience, we rewrite this constraint as $(\|Dp\|^2 - \Delta^2)/2 = 0$.

As we will discuss in more detail in a few lectures, we can solve the equality-constrained optimization problem using the method of Lagrange multipliers. We define the *Langrangian* for the optimization problem to be

$$L(p, \lambda) = \frac{1}{2}\|f(x_k) + f'(x_k)p\|^2 + \frac{\lambda}{2}\left(\|Dp\|^2 - \Delta^2\right).$$

The solution to the constrained optimization problem satisfies the critical point equation $\partial L/\partial p = 0$ and $\partial L/\partial \lambda = 0$. The equation $\partial L/\partial p = 0$ is the same as the Tikhonov-regularized least squares problem with regularization parameter λ. Whether λ is treated as a regularization parameter or a multiplier that enforces a constraint is thus simply a matter of perspective. Hence, we can consider the Levenberg-Marquardt method as minimizing the model $\|f(x_k) + f(x_k)p\|$ subject to the constraint $\|Dp\| \leq \Delta$, where a larger or smaller value of λ corresponds to a smaller or larger value of Δ. We think of the region $\|Dp\| \leq \Delta$ as the region where the Gauss-Newton model provides good guidance for optimization; that is, it is a region where we trust the model.

4 Trust regions

A *trust region* method for mininizing ϕ involves a *model* $\mu(p)$ that is supposed to approximate the decrease $\phi(x_k + p) - \phi(x_k)$ associated with taking a step p; and a *trust region*, often chosen to be a sphere $\|p\|^2 \leq \Delta$, where we believe the model to provide reasonable predictions. At each step of the method, we (approximately) minimize the model within the trust region to get a proposed step p, then check the gain ratio associated with taking that step:

$$\rho_k = \frac{\phi(x_k) - \phi(x_k + p_k)}{\mu(0) - \mu(p_k)}.$$

Depending on whether the gain ratio, we adjust Δ; a strategy proposed in Nocedal and Wright is:

- If $\rho_k < 1/4$, we were too aggressive; set $\Delta_{k+1} = \Delta_k/4$.
- If $\rho_k > 3/4$ and $\|p_k\| = \Delta_k$, we were too conservative; set $\Delta_{k+1} = \min(2\Delta_k, \Delta_{\max})$.
- Otherwise, leave $\Delta_{k+1} = \Delta_k$.

We also use the gain ratio to decide whether to accept or reject the step. For $\rho_k > \eta$ for a fixed $\eta \in [0, 1/4)$, we accept ($x_{k+1} = x_k + p$); otherwise we reject ($x_{k+1} = x_k$).

Compared to a line search strategy, trust region methods have the advantage that we adapt not just the step length but also the direction of the search. Consequently, trust region methods often exhibit more robust convergence, though both line search and trust region approaches exhibit good global convergence properties, and both approaches lead to eventual superlinear convergence when paired with a Newton model (i.e. a quadratic approximation centered at x_k) or a quasi-Newton method such as BFGS.

5 Inexact search and the dog-leg

One of the main difficulties with the trust region approach is solving a constrained quadratic optimization as a subproblem. Because we do not know the Lagrange multiplier in advance, solving this problem exactly requires several times the effort of solving an unconstrained problem, as we might do in an ordinary Newton or quasi-Newton method without the trust region modificiation. As with line search, though, the cost of doing an exact search is probably not worthwhile — we would rather get a good-enough approximate solution and move on.

A popular inexact search approach is the *dog leg* method[1]. The idea of the dog leg method is to approximate the shape of the curve

$$p(\Delta) = \operatorname{argmin}_p \mu(p) \text{ s.t. } \|p\| \leq \Delta$$

based on the observation that

- $p(0) = 0$.
- $p'(0) \propto -\nabla\phi(x_k)$.

[1] There is, in fact, a double dogleg method. I double dogleg dare you to look it up.

- For large Δ, $p(\Delta) = p_\infty$ is the unconstrained minimizer of μ.

We thus approximate the $\rho(\Delta)$ curve by a piecewise linear curve with

- A line segment from 0 to $-\alpha\nabla\phi(x_k)$ where $\mu(-\alpha\nabla\phi(x_k))$ is mimimized.
- Another line segment from $-\alpha\nabla\phi(x_k)$ to p_∞.

A related approach is *two-dimensional subspace minimization*, which involves a constrained miminization over the two-dimensional subspace spanned by $-\nabla\phi(x_k)$ and p_∞.

The *Steighaug* method combines the trust region approach with a (linear) conjugate gradient solve on the quadratic model problem. The idea is to trace out a polygonal path (as in the dog leg method) connecting the CG iterates, until that path intersects the trust region boundary. If the (approximate) Hessian used by the model is indefinite, CG runs until it discovers the indefiniteness, then plots a path toward where the model descends to $-\infty$. There are more recent variants which combine Newton, trust regions, and Krylov subspaces in various clever ways; other than mentioning that they exist, though, we leave this topic for the interested student to pursue in her copious free time.