1 Regularization

The quadratic approximation of the objective function $f(x)$ at iterate x_k is

$$\tilde{f}(x_k + \Delta x) = f(x_k) + \nabla f(x_k)'\Delta x + \frac{1}{2}\Delta x' H_k \Delta x,$$

where H_k is the Hessian $\nabla^2 f(x_k)$ or an approximation thereof. The search direction Δx from iterate x_k is computed by minimizing this function over the variable Δx resulting in the equation

$$H_k \Delta x = -\nabla f(x_k) := -y_k \tag{1}$$

for the search direction Δx. Clearly we can assume that $y_k = \nabla f(x_k) \neq 0$ since the search terminates at a zero gradient .

In our algorithms we can also assume that H_k is positive semidefinite, but not necessarily nonsingular. If H_k is nonsingular (i.e. positive definite), then the solution Δx is a *descent direction*:

$$\Delta x' \nabla f(x_k) = -\Delta x' H_k \Delta x < 0$$

and this is all we care about. We will solve (1) by Cholesky factorization $H_k = LL'$ with lower triangular L, by solving the triangular systems

$$Lv = -y_k, \quad L'\Delta x = v. \tag{2}$$

The Cholesky factorization fails if H_k is singular. In that case we replace H_k with $H_k(\delta) := H_k + \delta I$, for some positive constant δ. This matrix is now positiv definite, in fact

$$(H_k(\delta)u, u) = u' H_k(\delta) u = u' H_k u + \delta u' I u \geq \delta \|u\|^2$$

so that (1) now yields a descent direction. Moreover it improves the conditioning of (1): if $H_k + \delta I = L(\delta)L(\delta)'$ is the Cholesky factoriztion of $H_k(\delta)$, then we have

$$|L(\delta)_{ii}| \geq \delta.$$

Indeed, the diagonal element $L(\delta)_{ii}$ is an eigenvalue of the triangular matrix $L(\delta)'$. Let u be a corresponding eigenvector. Then we have

$$|L(\delta)_{ii}|^2 \|u\|^2 = \|L(\delta)'u\|^2 = (L(\delta)L(\delta)'u, u) = ((H_k(\delta)u, u) \geq \delta \|u\|^2$$

from which it follows that

$$|L(\delta)_{ii}| \geq \sqrt{\delta}$$

with obvious implications for the numerical stability of the triangular systems (2). Moreover this suggests that we should replace H_k with $H_k + \delta I$ not only if the Cholesky factorization fails but rahter as soon as the minimal diagonal element (in absolute value) of the Cholesky factor L is below the threshold $\sqrt{\delta}$.

Trust region interpretation. The passage from the matrix H_k to the regularization $H_k + \delta I$ has an interpretation in terms of *trust regions*: the solution Δx^* of

$$H_k(\delta)\Delta x = -y_k, \quad \text{where } y_k = \nabla f(x_k),$$

is the minimizer of the quadratic function

$$\begin{aligned}\phi(\Delta x) &= f(x_k) + y_k'\Delta x + \Delta x' H_k(\delta)\Delta x \\ &= f(x_k) + y_k'\Delta x + \Delta x' H_k \Delta x + \delta \|\Delta x\|^2 \\ &= \tilde{f}(x + \Delta x) + \delta \|\Delta x\|^2.\end{aligned}$$

Now note that this minimizer Δx^* is automatically also the minimizer of the quadratic approximation $\tilde{f}(x_k + \Delta x)$ on the ball $B(x_k, r_k)$ with radius $r_k = \|\Delta x^*\|$. Indeed, if this ball contained a point u with $\tilde{f}(x_k + u) < \tilde{f}(x_k + \Delta x^*)$, then, since also $\|u\| \leq \|\Delta x^*\|$ it follows that

$$\phi(u) = \tilde{f}(x_k + u) + \delta \|u\|^2 < \tilde{f}(x_k + \Delta x^*) + \delta \|\Delta x^*\|^2 = \phi(\Delta x^*).$$

In other words: passing from H_k to $H_k + \delta I$ we compute the search direction Δx by minimizing the quadratic approximation $\tilde{f}(x_k + \Delta x)$ not globally but instead on the ball $B(x_k, r_k)$ (the region in which we trust the approximation) where the trust radius r_k is defined implicitly as $r_k = \|\Delta x^*\|$.

This indicates that the regularization $H_k \to H_k(\delta)$ is not unreasonable and in any case it solves the problem of nonsingularity of H_k for us, improves the conditioning and results in a descent direction Δx at iterate x_k.

2 Hessian

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. The second order Taylor expansion of f centered at x has the form

$$f(x + h) = f(x) + L(h) + \tfrac{1}{2}B(h, h) + R(h),$$

where L is a linear function of $h \in \mathbb{R}^n$, $B(u, v)$ is a bilinear function of $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n$ and the remainder $R(h)$ satisfies $R(h) = o(\|h\|^2)$. This condition on the remainder ensures that L and B are uniquely determined as

$$L(h) = \nabla f(x)'h \quad \text{and} \quad B(u, v) = u'Hv,$$

where $H := \nabla^2 f(x) \in Mat_{n\times n}(\mathbb{R})$ is the matrix with entries

$$H_{ij} = B(e_i, e_j) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x),$$

i.e. the Hessian matrix of f at x.

To compute the gradient and Hessian of a C^2-function f we make use of the fact that the remaider condition $R(h) = o(\|h\|^2)$ in a quadratic expansion

$$f(x + h) = f(x) + \Delta' h + h' H h + R(h)$$

uniquely determines the "coefficients" Δ and H as $\Delta = \nabla f(x)$ and $H = \nabla^2 f(x)$. We only need to find such an expansion for $f(x+h)$ and check that the remainder satisfies $R(h) = o(\|h\|^2)$. This is how we will derive our formulas below.

Hessian of affine transformation. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^2-function and $\overline{f}(u) = f(x_0 + Fu)$, where $x_0 \in \mathbb{R}^n$ and $F : \mathbb{R}^m \to \mathbb{R}^n$ is a linear map, i.e. $F \in Mat_{n\times m}$.

We want to compute the gradient and Hessian of h at any point u from those of f. To get these do a second order Taylor expansion of f about x:

$$f(x+h) = f(x) + h^T g + h^T Hh + o(\|h\|^2),$$

where $g = \nabla f(x)$ and $H = \nabla^2 f(x)$ are uniquely determined by the fact that the residual is $o(\|h\|^2)$. Applying this to the point $x = x_0 + Fu$ this implies that

$$\begin{aligned}\overline{f}(u+h) &= f(x_0 + Fu + Fh) \\ &= f(x_0 + Fu) + (Fh)^T g + (Fh)^T HFh + o(\|Fh\|^2).\end{aligned}$$

Since $o(\|Fh\|^2)$ is $o(\|h\|^2)$ we conclude from this that

$$\nabla\overline{f}(u) = F^T g \quad \text{and} \quad \nabla^2\overline{f}(u) = F^T HF,$$

or, more explicitly

$$\nabla\overline{f}(u) = F^T \nabla f(x_0 + Fu) \quad \text{and} \quad \nabla^2\overline{f}(u) = F^T \nabla^2 f(x_0 + Fu)F. \tag{3}$$

Hessian of composition. With a similar approach we can compute the Hessian of a composition $g(f(x))$, where here $g : \mathbb{R} \to \mathbb{R}$ is a scalar function of one variable (more general g are much harder to handle and we do not need them). Indeed, set

$$y = f(x), \quad \nabla = \nabla f(x) \quad H = \nabla^2 f(x) \quad \text{and} \quad k = h^T\nabla + \tfrac{1}{2}h^T Hh$$

and use second order Taylorapproximations on f at the point x and g at the point $y = f(x)$ to obtain:

$$\begin{aligned}g(f(x+h))) &= g\left(f(x) + h^T\nabla + \tfrac{1}{2}h^T Hh\right) \\ &= g(y+k) = g(y) + g'(y) + \tfrac{1}{2}g''(y)k^2 + o(k^2) \\ &= g(y) + g'(y)\left(h^T\nabla + \tfrac{1}{2}h^T Hh\right) + \frac{1}{2}g''(y)\left(h^T\nabla + \tfrac{1}{2}h^T Hh\right)^2 + o(\|h\|^2).\end{aligned}$$

Here we have used that $o(k^2) = o(\|h\|^2)$. Collect terms of first and second order in h together and sticking all terms of higher order into the residual $o(\|h\|^2)$. Note that the squared term contributes no first order terms and only one second order term, this being the term

$$(h^T\nabla)^2 = (h^T\nabla)(h^T\nabla) = (h^T\nabla)(h^T\nabla)^T = h^T(\nabla\nabla^T)h.$$

We obtain

$$g(f(x+h))) = g(y) + h^T[g'(y)\nabla] + \frac{1}{2}h^T\left[g'(y)H + g''(y)\nabla\nabla^T\right]h + o(\|h\|^2).$$

and from this we can read off that

$$\nabla(g\circ f)(x) = g'(f(x))\nabla f(x), \quad \text{and} \tag{4}$$

$$H(g\circ f)(x) = g'(f(x))H + g''(f(x))\nabla f(x)\nabla f(x)^T \tag{5}$$

Note that here $d = \nabla f(x)$ is viewed as a column vector and so $\nabla f(x)\nabla f(x)^T$ is the *outer product*

$$\nabla f(x)\nabla f(x)^T = dd^T = (d_i d_j)_{ij}.$$

We will need the following example to construct test functions for unconstrained minimization:

Example 2.1. Let $\phi : G \subseteq \mathbb{R} \to \mathbb{R}$ be a function of one variable defined on an open subset $G \subseteq \mathbb{R}$, fix $a \in \mathbb{R}^n$ and set $g(x) = \phi(a \cdot x)$, for all $x \in \mathbb{R}^n$ such that $a \cdot x \in G$.

Since $f(x) = a \cdot x$ satisfies $\nabla f(x) = a$ and $H = \nabla^2 f(x) = 0$, for all $x \in \mathbb{R}^n$, the formulas (4) and (5) yield

$$\nabla g(x) = \phi'(a\cdot x)a \quad \text{and} \quad \nabla^2 g(x) = \phi''(a\cdot x)aa'.$$

The idea is to construct test functions of the form

$$objF(x) = \sum\nolimits_j \phi_j(a_j \cdot x)$$

where all the $\phi_j = \phi_j(u)$ have a unique minimum at $u = 0$. Then $objF(x)$ assumes its global minimum at all points x satisfying $a_j \cdot x = 0$, for all j, equivalently $Ax = 0$, where A is the matrix with rows a_j, provided such a solution exists.

If the functions ϕ_j are all convex, the same is true of our objective function $objF$ (sums and compositions of convex functions are again convex) With this we can construct examples with well conditioned, poorly conditioned and even singular Hessians ($ker(A) \neq \{0\}$) with known minimizers. The conditioning of $\nabla^2 f(x)$ is closely related to that of the matrix A.