Notes on Nocedal and Wright's "Numerical Optimization"
Chapter 6 – "Calculating Derivatives"

Lucas N. Ribeiro

Introduction I

- Sometimes the user can provide a function that calculates the gradient and the Hessian; sometimes not.
- When these information are not available, we can calculate them by ourselves
- There are basically three approaches to achieve that:
  - Finite differencing: Observe the response of the function of interest to small perturbations, e.g., using the central difference formula;
  - Automatic differentiation: breaks the function down into elementary operations and applies the chain rule.
  - Symbolic differentiation: algebraic specification and symbolic manipulation. Used in Maple, MATLAB, etc.
- Besides algorithms, gradients and Hessians are useful for sensitivity analysis in economics, etc.

Finite-Difference Derivative Approximations I

Let us first derive the forward-difference formula from Taylor's theorem. If f is twice continuously differentiable, then:

$$f(x+p) = f(x) + \nabla f(x)^{\mathsf{T}} p + 0.5 p^{\mathsf{T}} \nabla^2 f(x+tp) p \tag{1}$$

Let us assume L is a bound on the norm of $\nabla^2 f(\cdot)$. It means that

$$\exists L \mid \|\nabla^2 f\| < L$$

for whatever argument of the Hessian. Then, the norm of the last term in (1) is bounded as

$$\begin{aligned} \|0.5 p^{\mathsf{T}} \nabla^2 f(x+tp) p\| &< 0.5 \|p^{\mathsf{T}}\| \|\nabla^2 f(x+tp)\| \|p\| && (2) \\ &< (L/2) \|p\|^2 && (3) \end{aligned}$$

Finite-Difference Derivative Approximations II

where we used Cauchy-Schwarz. By isolating the quadratic term in (1) and applying its bound, we get:

$$\|f(x+p) - f(x) - \nabla f(x)^{\mathsf{T}} p\| \leq (L/2)\|p\|^2 \tag{4}$$

Now, set $p = \epsilon e_i$, where e_i is the ith canonical vector. Then

$$\nabla f(x)^{\mathsf{T}} p = \epsilon \partial f / \partial x_i$$

Equation (4) can be rearranged as

$$\|f(x+\epsilon e_i) - f(x) - \epsilon \partial f / \partial x_i\| \leq (L/2)\epsilon^2 \tag{5}$$

The inequality above implies

$$\partial f / \partial x_i = \frac{f(x+\epsilon e_i) - f(x)}{\epsilon} + \delta \tag{6}$$

Finite-Difference Derivative Approximations III

where $|\delta| \leq (L/2)\epsilon$. When $\epsilon \to 0$, the error vanishes and the finite difference goes to the partial derivative.
A typical choice for ϵ is $\sqrt{\mathbf{u}}$, where $\mathbf{u}$ is the unit round-off.
A more accurate approximation is the central difference formula.
Assuming the existence of second order derivative of f and Lipschitz continuity:

$$
\begin{aligned}
f(x+p) &= f(x) + \nabla f(x)^{\mathsf{T}} p + 0.5 p^{\mathsf{T}} \nabla^2 f(x+tp) p && (7) \\
&= f(x) + \nabla f(x)^{\mathsf{T}} p + 0.5 p^{\mathsf{T}} \nabla^2 f(x) p + O(\|p\|^3). && (8)
\end{aligned}
$$

We set $p = \epsilon e_i$ and $p = -\epsilon e_i$. Then we have:

$$
f(x+\epsilon e_i) = f(x) + \epsilon \frac{\partial f}{\partial x_i} + 0.5\epsilon^2 \frac{\partial f^2}{\partial x_i^2} + O(\epsilon^3) \qquad (9)
$$

$$
f(x-\epsilon e_i) = f(x) - \epsilon \frac{\partial f}{\partial x_i} + 0.5\epsilon^2 \frac{\partial f^2}{\partial x_i^2} + O(\epsilon^3) \qquad (10)
$$

Finite-Difference Derivative Approximations IV

Subtracting (10) from (9) and dividing by 2ϵ gives:

$$\frac{\partial f}{\partial x_i} = \frac{f(x + \epsilon e_i) - f(x - \epsilon e_i)}{2\epsilon} + O(\epsilon^2) \tag{11}$$

- Now the error is $O(\epsilon^2)$ (unlike $O(\epsilon)$ in the forward-difference)
- More complex, though. Has to evaluate f at two different points.

Approximating a Jacobian – Now consider a function $r : \mathbb{R}^n \to \mathbb{R}^m$. The Jacobian matrix is defined as $J(x) = [\partial r_m / \partial x_n]$. Using Taylor's theorem, we can deduce that

$$\|r(x + p) - r(x) - J(x)p\| \leq (L/2)\|p\|^2$$

Finite-Difference Derivative Approximations V

Sometimes we wish an estimate for $J(x)p$ instead of the full Jacobian. In this case, we have

$$J(x)p = \frac{r(x + \epsilon p) - r(x)}{\epsilon} + O(\epsilon)$$

When we wish the full matrix, we can compute one column at a time:

$$\partial r / \partial x_i = \frac{r(x + \epsilon e_i) - r(x)}{\epsilon} + O(\epsilon)$$

Computationally efficient methods that exploit possible sparsity can be derived.

Approximating the Hessian – From Taylor's theorem:

$$\nabla f(x + p) = \nabla f(x) + \nabla^2 f(x) p + O(\|p\|^2)$$

and

$$\partial \nabla f / \partial x_i = \frac{\nabla f(x + \epsilon e_i) - \nabla f(x)}{\epsilon}$$

Finite-Difference Derivative Approximations VI

Sometimes we wish to estimate only $\nabla^2 f(x)p$ and we can achieve it by

$$\nabla^2 f(x)p \approx \frac{\nabla f(x+\epsilon p) - \nabla f(x)}{\epsilon}$$

Note that, so far, we require knowledge of the gradient vector. When it's not available we can plug our gradient approximations into that of Hessian:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{f(x+\epsilon e_i+\epsilon e_j) - f(x+\epsilon e_i) - f(x+\epsilon e_j) + f(x)}{\epsilon^2} + O(\epsilon). \tag{12}$$

Automatic Differentiation

- Any function (no matter how complicated) is built by simple operation such as sums, multiplications and exponentiations.
- The basic principle of automatic differentiation is breaking down functions in these operations and applying the chain rule.