A brief overview of practical optimization methods

January 14, 2010

Antoine Goujard[1]

Refererences:

- Handout 8. Numerical optimization, provides a lot of examples of numerical optimization and shows the computational costs associated with each method for many test functions with various shapes.
- Train, 2009, Discrete Choice Methods with Simulation. Chapter 8, `http://elsa.berkeley.edu/books/choice2nd/Ch08_p183-204.pdf`. Train summarizes in an efficient way the main derivative based methods (Newton Raphson, Berndt-Hall-Hall-Hausman, steepest ascent and their step-adjusted versions: Davidson-Flechter-Powell and Broyden-Flechter-Goldfarb-Shanno). However this chapter does not have any information on the non-derivative(=non-gradient) based methods.
- Cameron and Trivedi (2005), Microeconometrics Methods and Applications. Chapter 10 is brief 16p and provides an overview of the simulated annealing algorithm, but no information on the Nelder-Mead algorithm(=downhill-simplex=amoeaba).
- Cameron and Trivedi (2009), Microeconometrics Using Stata. Chapter 11 review the maximum likelihood estimation and optimiziation commands in stata and its associated matrix language, mata.

Here we want to solve a maximization (or minimization) problem:
$\theta^* = argmax_{\theta \in \Theta} F(\theta)$ where F maps $\mathbb{R}^k$ to $\mathbb{R}$.[2]
However, due to the particular structure of this problem there may be no explicit/closed form solution(s) to this problem. A particular case with an explicit solution is the least squares problem:
$b^* = argmax_b ||y - X.b||^2$ where b is in $\mathbb{R}^k$ and $||.||$ is the euclidian norm in $\mathbb{R}^n$. When $rank(X) = k \leq n$ the solution is unique and $b^* = (X'X)^{-1}X'y$.

In general optimization problems do not have explicit solutions and we have to rely on some set of necessary/sufficient conditions to define the optimal value. For example, when we try to maximize a Log-likelihood, the theory

[1]Some of you asked me to make this summary. Please let me know if you find any typo or mistake, a.j.goujard@lse.ac.uk. Note that this is unofficial material and use at your own risk.

[2]Note that in Econometrics F depends on the particular sample considered and that θ^* will be an estimate of the true parameter value. Here I depart from the notations F_n and $\hat{\theta}_n$ that emphasize these relationships.

tells us that θ^* has to satisfy some first order and second order conditions (FOC and SOC), but we are often unable to write down an explicit expression G, such that: $\theta^* = G(y, X)$. If we focus on the FOCs, they define a functional equation in θ^* that has to be solved numerically and may have several solutions. The **critical** points are the zeros of the **gradient** vector, ie $\frac{\partial F}{\partial \theta}(\theta^*) = 0_k$. However, a critical point may not be a global maximum or even a maximum. We may reach **a local maximum, a minimum or a saddle point**. If the last two cases can be ruled out using the Hessian matrix, the first problem remains. We need to put more structure on our initial problem. For example, if $F()$ is **globally and strictly concave** our problem should have a unique solution, which, if the function is differentiable, corresponds to the FOCs.

In econometric practice, $F()$ is often not so well-behaved and the optimized functions have very different shapes and characteristics. As a result, the appropriate numerical method(s) will also change. However, most methods share the same **iterative** structure. They iter an initial guess about θ^*, the initial value $\theta^{(0)}$ until the value $\theta^{(m)}$ appears to satisfy some defined criterions or stopping rules. Two main features determine the **efficiency** of these algorithms:

- The **shape of the function** $F()$. Important features are global concavity, the absence of area such that $F()$ is nearly flat [3], the fact that you are able to differentiate $F()$ to find analytical expression for the gradient or the Hessian matrix.

    1. If the function is not differentiable or even discontinuous[4], it is not possible to rely on local approximation of F such has in Newton-Raphson and other derivative methods.
    2. The presence of local maxima or other critical points will also influence the convergence of the algorithms towards the global maximum.

- The **computational costs of the method**

    1. The number of iterations (m) to obtain a satisfying value $\theta^{(m)}$.
    2. The number of evaluations of the objective function at each step.
    3. The other functions to evaluate at each step (eg. The gradient and the Hessian for Newton-Raphson) and the easiness to compute them (do these functions have an analytical form or do we have to rely on a numerical approximation?).

[3] The Rosenbrock's function (p95 H8) is the classical example of a nearly flat function.
[4] This is the case for the Maximum score estimator of Manski.

These computations may cause some **practical problems**. **Rounding errors** and other approximations may prevent the convergence of an algorithm even if it should converge to a global maximum [5].

In the problem sets, we have mainly seen two iterative methods(see table 1, H8, p94): Newton Raphson (NR) and Berndt-Hall-Hall-Hausman (BHHH). They are based on the following iterations:
$\theta^{(m)} = \theta^{(m-1)} + S^{(m)}$
With $S^{(m)} = -H^{-1}(\theta^{(m-1)}).\frac{\partial F}{\partial \theta}(\theta^{(m-1)})$ for NR.
And $S^{(m)} = [\frac{\partial F}{\partial \theta}(\theta^{(m-1)})\frac{\partial F}{\partial \theta}(\theta^{(m-1)})^t]^{-1}.\frac{\partial F}{\partial \theta}(\theta^{(m-1)})$ for BHHH.
To compare the 2 methods, let's see their requirements: (1) in terms of functional shape F. To compute the steps we need:

- F, the gradient and the Hessian of F at $\theta^{(m-1)}$ for NR. Moreover, if we want the steps, $S^{(m)}$, to go into a direction of improvement of F, we need H to be full-rank and negative-semi definite (ie. negative definite) at each iteration m and not only at the local maxima. This will be the case for a globally strictly concave function.

- F, the gradient of F at $\theta^{(m-1)}$ for BHHH. Moreover, the steps always go into a direction of improvement of F, because $[\frac{\partial F}{\partial \theta}(\theta^{(m-1)})\frac{\partial F}{\partial \theta}(\theta^{(m-1)})^t]$ is positive definite.

Then, (2) we may be interested in the efficiency of the two methods. For a well behaved quadratic function, NR converges in one step. Indeed, if we have:
$F(\theta) = a + B.\theta + \theta^t.C.\theta$ with C a negative definite matrix.
Then $S^{(1)} = -1/2.C^{-1}.(B^t + 2.C.\theta^{(0)})$ so:
$\theta^{(0)} + S^{(1)} = -1/2.C^{-1}.B^t + (I - C^{-1}.C)\theta^{(0)} = -1/2.C^{-1}.B^t = \theta^*$. This shows that NR converges **quadratically**. Moreover, we expect this method to works well if a quadratic approximation of F is accurate or if the initial value, $\theta^{(0)}$, is in a neighborhood of θ^*. However the NR method is **computationally intensive** as it requires to compute and invert the Hessian matrix at each step. Here, BHHH is less demanding as it requires only to compute the gradient of F. But, BHHH can be seen as an approximation of NR[6] and this approximation can be quite bad, especially if $\theta^{(0)}$ is far from the optimal value θ^*.

Once we are able to compute the steps, we need to check that the final value $\theta^{(m)}$ is indeed a **global maximum**. This may fail for both NR and BHHH.

[5] This is the case in PS5, question 1.b. We tried to optimize the Poisson likelihood function which is globally concave. The derivative based method, GRADX, should converge but appears to fail for some initial values.

[6] The approximation is based on the information matrix identity if we are trying to maximize a log-likelihood.

The stopping rules (convergence test), may lead to a local critical point. It is important to check the Hessian, try several initial values $\theta^{(0)}$ and test for a global minimum (Veall's test, p94 H8).