STAT 538 Homework 2
Out January 17, 2011
Due January 26, 2011
©Marina Meilă
mmp@stat.washington.edu

Submit the code used to solve these problems through the `Assignments` web page. Turn in the required solutions (without the code) in class on the due date. Only the paper part of the homework is graded.

In general, I require you to write your own code. For this assignment however, there will be some exceptions, noted in the text.

Problem 1 – Comparison of unconstrained minimization methods. A function with non-convex level sets[1]

The task is to find numerically the minimum of the following function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, f(x) = x_2^2 - a x_2 ||x||^2 + ||x||^4.$$

for $a = 1.98$. In the above $x = [x_1\, x_2]^T$ is a two-dimensional vector. Start from the point $x^0 = (0.5, 0.5)$.

1. Write the formulas of the gradient and Hessian of f.

2. Find $\min_x f$ using the following methods:

Gconst Steepest descent with constant step size. Find a suitable step size by e.g trial and error.

Gsearch Steepest descent with line search. Use either one of: the Brent method, the Golden section algorithm, or te Armijo rule. (OK to transcribe or to download source code for line minimization.)

N Newton with line search.

QN Optional, for extra credit and highly encouraged: Quasi-Newton (OK to use source code for the Quasi-Newton step only; a search for `L-BFGS code` returns various downloadable sources)

For each of the algorithms above, plot:

- the values $f(x^k)$ versus k (consider a logarithmic plot)

[1] This problem is problem 1.1.10 from B

- the search path $(x^k)_{k=0,1,...}$ in 2D
- anything else that you find interesting.

You are strongly recommended to plot the results of all methods on the same graph for better comparison (i.e all f plots on one graph, all x^k plots on another graph, etc)

Record also: the final values of x, f and the number of iterations to convergence for each method.

Use the stopping criterion $||(\nabla^2 f(x^k))^{-1}\nabla f(x^k)|| \leq \epsilon = 10^{-5}$ for all experiments.

Note: A simple calculation shows that $f(x) > 0$ for $x \neq 0$, hence the global minimum is at $x^* = 0$. Notation $||x||^2 = x_1^2 + x_2^2$. If you "cheat" changing the x^0 you can make the problem much easier. Try this, but you will receive full credit only if you start from the required x^0.

Unlike the next problem where f is convex, this function is not, and its level sets are not convex near the origin. It is an example of an artificially constructed "hard" function to benchmark optimization algorithm.

Problem 2 – Maximum Likelihood estimation of the logistic density parameters

The logistic density is defined by

$$p(x; a, b) \;=\; \frac{ae^{-ax-b}}{(1+e^{-ax-b})^2} \quad \text{for } x \in \mathbb{R}$$

For a given data set $\mathcal{D}_N$ containing N i.i.d. points sampled from p, define the function f to be $f(a,b) = -\frac{1}{N}\ln p(\mathcal{D}; a, b)$. This problem is about the numerical estimation of the Maximum Likelihood parameters of the logistic distribution. You will use the data in the file `hw2-logistic-data.dat` available on the `Assignmets` web page. The file contains $N = 2000$ real values, in plain text format, one per line.

1. Write the formulas of the gradient and Hessian of f, i.e $\frac{\partial f}{\partial a}$, $\frac{\partial f}{\partial b}$, $\frac{\partial^2 f}{\partial a \partial b}$, $\frac{\partial^2 f}{\partial a^2}$, $\frac{\partial^2 f}{\partial b^2}$.

2. Find the parameters a, b by minimizing the function f using the following methods:

1. Steepest descent with constant step size. Find a suitable step size by e.g trial and error.

2. Steepest descent with line search. Use either one of: the Brent method, the Golden section algorithm, or te Armijo rule with bracketing the minimum. (OK to transcribe or to download source code for line minimization.)
3. Newton with line search.
4. Stochastic gradient (with diminishing step size). Use $\lambda = 1/4$.

   [Optional, extra credit, hard] Assume that $b = 0, a$ known. Show how to choose λ in this case. Simplify your work by replacing the sample with the original p, and therefore averaging over samples with expectations under p. Some approximations may be necessary.
5. Optional, for extra credit and highly encouraged: Quasi-Newton (OK to use source code for the Quasi-Newton step only; a search for `L-BFGS code` returns various downloadable sources)

For each of the algorithms above, plot:

- the values $f(a^k, b^k)$ versus k
- the search path $(a^k, b^k)_{k=0,1,...}$ in 2D
- anything else that you find interesting.

For the plots you submit, start all the algorithms from the same initial point $a = 1$, $b = 0$. You are strongly recommended to plot the results of all methods on the same graph for better comparison (i.e all f plots on one graph, all $a(^k, b^k)$ plots on another graph, etc)

Record also: the final values of $a, b, f(a, b)$ and the number of iterations to convergence for each method.

Optional, but highly recommended, as it helps you test your result: plot the data and fitted density $p(x; a, b)$ on the same graph. You are also encouraged to try different starting points and other variations to the methods and to notice the differences between outcomes without necessary submitting the work for grading.

Use the stopping criterion $||(\nabla^2 f(x^k))^{-1} \nabla f(x^k)|| \leq \epsilon = 10^{-5}$ for all experiments.

[**Optional** Display the level sets of f on the plot of (a^k, b^k).]

Some helping remarks

- Note that the inverse of a *symmetric* 2×2 matrix H is given by

$$H^{-1} = \frac{1}{h_{11}h_{22} - h_{12}^2} \begin{bmatrix} h_{22} & -h_{12} \\ -h_{12} & h_{11} \end{bmatrix} \tag{1}$$

- The logistic density is only defined for $a > 0$. So, theoretically, this is a constrained problem. However, all the algorithms you implement and run, unless you have bugs or bad step sizes, should have no problem avoiding the "forbidden" zone. (Can you see why?)