Continuous Maths HT 2019: Problem Sheet 4

Advanced Root-Finding and Numerical Optimization

4.1 Suppose that $f : \mathbb{R} \to \mathbb{R}$ has three continuous derivatives, and a *double root*: $f(x^*) = \frac{\mathrm{d}f}{\mathrm{d}x}(x^*) = 0$ and $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x^*) \neq 0$. We can find it using the so-called *relaxed* Newton iteration $x_{n+1} = x_n - \alpha \frac{f(x_n)}{\frac{\mathrm{d}f}{\mathrm{d}x}(x_n)}$, where α is a constant.

(a) Show that
$$\epsilon_{n+1} = \frac{\epsilon_n \frac{\mathrm{d}f}{\mathrm{d}x}(x_n) - \alpha f(x_n)}{\frac{\mathrm{d}f}{\mathrm{d}x}(x_n)}.$$

(b) In the numerator use Taylor's theorem for $\frac{\mathrm{d}f}{\mathrm{d}x}(x_n)$, at x^* with second-order remainder term, and Taylor's theorem for $f(x_n)$, at x^* with third-order remainder term. In the denominator use Taylor's theorem for $\frac{\mathrm{d}f}{\mathrm{d}x}(x_n)$, at x^* with first-order remainder term. Show that for a certain choice of α, $|\epsilon_{n+1}| \leq A|\epsilon_n|^2$, for some A which is finite when x_n is sufficiently close to x^*.

(c) Deduce that, if x_0 is close to x^*, the relaxed iteration converges quadratically.

(d) Optional: what about a root of order m, where $\frac{\mathrm{d}^n f}{\mathrm{d}x^n}(x^*) = 0$ for $n < m$ and $\frac{\mathrm{d}^m f}{\mathrm{d}x^m}(x^*) \neq 0$?

4.2 If $Y \sim \mathrm{Geo}(p)$, i.e. $\mathrm{P}[Y = k] = (1-p)^k p$ for $k \geq 0$, recall that the p.g.f. of Y is $G_Y(s) = \frac{p}{1-(1-p)s}$. *If you do not recall this, quickly derive it for yourself.*

(a) In a two-type branching process like **Example 5.7**, let Y_{AA} be the number of type-A offspring of an individual of type A, Y_{AB} the number of type-B offspring of an individual of type A, and so on. Suppose that
$$Y_{AA} \sim \mathrm{Geo}(\tfrac{1}{2}), \quad Y_{AB} \sim \mathrm{Geo}(\tfrac{1}{2}), \quad Y_{BA} \sim \mathrm{Geo}(\tfrac{1}{3}), \quad Y_{BB} \sim \mathrm{Geo}(\tfrac{2}{3}).$$
Show that the Jacobian for the root-finding problem in **Example 5.7**, which determines the probability of extinction after starting with one type-A individual, x, and after starting with one type-B individual, y, is
$$\begin{pmatrix} \frac{1}{(2-x)^2(2-y)} - 1 & \frac{1}{(2-x)(2-y)^2} \\ \frac{4}{(3-2x)^2(3-y)} & \frac{2}{(3-2x)(3-y)^2} - 1 \end{pmatrix}.$$

(b) Write a program to implement Newton's method in two dimensions, and use it to find the extinction probabilities for this two-type branching process. If you wish, explore how the parameters to the geometric distributions affect whether extinction is certain.

4.3 Prove the *Sherman-Morrison formula*: if $\mathbf{A}$ is an invertible $n \times n$ matrix, $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, and $\mathbf{A} + \boldsymbol{u}\boldsymbol{v}^T$ is invertible, then
$$(\mathbf{A} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\mathbf{A}^{-1}}{1 + \boldsymbol{v}^T\mathbf{A}^{-1}\boldsymbol{u}}.$$
Hint: in your calculation, keep an eye open for scalars.

Use this in Broyden's method to find $\hat{\mathbf{J}}_{\boldsymbol{n}}^{-1}$ in terms of $\hat{\mathbf{J}}_{\boldsymbol{n-1}}^{-1}$, $\boldsymbol{\Delta x}$, and $\boldsymbol{\Delta y}$. Using big-O notation give the time complexity of an efficient implementation of the iterative step, in terms of the dimension d, explaining your answer.

4.4 Consider the following improvement to golden section search, called *successive parabolic interpolation*: given a bracket (a,b,c), find the quadratic function $\hat{f}$ which interpolates $\big(a,f(a)\big)$, $\big(b,f(b)\big)$, $\big(c,f(c)\big)$; set z to be the minimum of this parabola; if $f(b)<f(z)$ the new bracket is (a,b,z), otherwise it is (b,z,c). Repeat until the bracket is small enough.

Derive the details of this method. How many ways can you find in which it might fail?

4.5 Consider the problem of finding $\arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$, where the Hessian of f is positive definite everywhere. Generally, gradient descent has linear convergence, and the ratio of successive errors depends on the *condition number* $\kappa\big(\mathbf{H}(f)\big)=\frac{\lambda_{\max}}{\lambda_{\min}}$ where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\mathbf{H}(f)$. Lower condition numbers give faster convergence.

Speed of convergence can often be improved by a linear change of variables: for a symmetric positive definite matrix $\mathbf{M}$, use gradient descent for $\arg\min_{\boldsymbol{y}} f(\mathbf{M}\boldsymbol{y})$ to generate the sequence $(\boldsymbol{y_0},\boldsymbol{y_1},\ldots,\boldsymbol{y_n})$, then recover $\boldsymbol{x_n}=\mathbf{M}\boldsymbol{y_n}$.

(a) Explain why $\frac{\mathrm{d}f(\mathbf{M}\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}(\boldsymbol{y_n})=\mathbf{M}\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}(\boldsymbol{x_n})$, and show that $\boldsymbol{x_n}$ can be computed directly, without any need to find the sequence $\boldsymbol{y_n}$, using the so-called *preconditioned* iteration

$$\boldsymbol{x_{n+1}}=\boldsymbol{x_n}-\alpha_n\mathbf{M}^2\tfrac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}(\boldsymbol{x_n}). \tag{$*$}$$

(b) Find the Hessian of $f(\mathbf{M}\boldsymbol{y})$ at $\boldsymbol{y_n}$ in terms of $\mathbf{H}(f)(\boldsymbol{x_n})$.

(c) Find a choice of $\mathbf{M}$ that leads to the lowest possible condition number for the first step of the iteration $(*)$. Explain the connection with Newton's method.

Hint: Since $\mathbf{H}(f)(\boldsymbol{x_0})$ is positive definite, it can be written $\boldsymbol{Q}^{-1}\boldsymbol{\Delta}^2\boldsymbol{Q}$, where $\boldsymbol{\Delta}$ is diagonal.

4.6 A method for 'loosely optimizing' the step size α_n is as follows. Given a direction $\boldsymbol{d}$ and preliminary choice of step size α', find a quadratic function $g(\alpha)$ that (i) interpolates $\big(0,f(\boldsymbol{x_n})\big)$ and $\big(\alpha',f(\boldsymbol{x_n}+\alpha'\boldsymbol{d})\big)$, and (ii) satisfies $\frac{\mathrm{d}g}{\mathrm{d}\alpha}(0)=\frac{\mathrm{d}f(\boldsymbol{x_n}+\alpha\boldsymbol{d})}{\mathrm{d}\alpha}(0)$. Then α_n is the minimum of g.

Why is this a good choice? What should be the conditions on $\boldsymbol{d}$ and/or α'? Find a formula for α_n.

4.7 If $\boldsymbol{f}:\mathbb{R}^d\to\mathbb{R}^m$, with $m>d$, then finding the $\boldsymbol{w}\in\mathbb{R}^d$ such that $\boldsymbol{f}(\boldsymbol{w})$ is as close as possible to $\mathbf{0}$ is called an *overdetermined system.* This means finding $\arg\min_{\boldsymbol{w}} l(\boldsymbol{w})$ where $l(\boldsymbol{w})=\|\boldsymbol{f}(\boldsymbol{w})\|^2=\boldsymbol{f}(\boldsymbol{w})^T\boldsymbol{f}(\boldsymbol{w})$.

(a) Express $\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{w}}$ in terms of $\mathbf{J}(\boldsymbol{f})$. Approximating $\mathbf{H}(l)$ by $2\mathbf{J}(\boldsymbol{f})^T\mathbf{J}(\boldsymbol{f})$ (challenge problem: justify this approximation), write down the quasi-Newton iterative step for the overdetermined system, where the step length is fixed to 1. (This is called the *Gauss-Newton method.*)

In the case of *linear regression*, $\boldsymbol{f}(\boldsymbol{w})=\mathbf{X}\boldsymbol{w}-\boldsymbol{y}$, where $\mathbf{X}$ is a $m\times d$ matrix and $\boldsymbol{y}\in\mathbb{R}^m$. This overdetermined system can be solved exactly using linear algebra.

(b) Derive an algebraic solution. Under what circumstances is your answer well-defined? How do you know that your answer is a minimum (as opposed to any other kind of stationary point)?

(c) Under the further constraint that $\|\boldsymbol{w}\|\le 1$, the system always has a well-defined answer. (This is called *ridge regression.*) Find $\boldsymbol{w}$ in terms of $\mathbf{X}$, $\boldsymbol{y}$, and a Lagrange multiplier μ, and explain why the formula is well-defined.