SciComp Practice Exam 5/19/14

*Answer the following questions explaining all steps that lead to a solution. Results presented without motivation will **not** receive any credit.*

1. Construct a fourth-order accurate approximation formula of $f'(x_0)$, based on values of $f\colon \mathbb{R} \to \mathbb{R}$ at points $x_i = x_0 + ih$, $i \in \mathbb{Z}$. Provide an estimate of the step size h that provides smallest relative error of the approximation in double precision floating point computation. Assume $f \in C^\infty(\mathbb{R})$.

Solution. Introduce notation $f_i = f(x_i)$, $f_i' = f'(x_i)$, $f_i^{(k)} = f^{(k)}(x_i)$ for $k > 1$. Problem asks for a fourth-order approximation $\mathcal{A}$, i.e.

$$f_0' = \mathcal{A}(f_{-p}, \ldots, f_q) + \mathcal{O}(h^4), p, q \in \mathbb{Z}.$$

Assume $\mathcal{A}$ linear, and note that Taylor series expansion around x_0 gives

$$\frac{1}{2h}(f_1 - f_{-1}) = f_0' + f_0^{(3)} \frac{h^2}{3!} + f_0^{(5)} \frac{h^4}{5!} + \ldots \tag{1}$$

$$\frac{1}{4h}(f_2 - f_{-2}) = f_0' + f_0^{(3)} \frac{(2h)^2}{3!} + f_0^{(5)} \frac{(2h)^4}{5!} + \ldots \tag{2}$$

Construct linear combination $\frac{4}{3} \times (1) - \frac{1}{3} \times (2)$ of above formulas to cancel $\mathcal{O}(h^2)$ term and obtain

$$f_0' = \frac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{12h} - \frac{4h^4}{5!} f_0^{(5)} + \mathcal{O}(h^4). \tag{3}$$

In floating point computations, subtraction of like quantities in the expression $-f_2 + 8f_1 - 8f_{-1} + f_{-2}$ leads to catastrophic loss of significant digits in approximation of f_0' as $h \to 0$. The absolute condition number (not relative in this case, since we're interested in behavior as $h \sim 0$) of the problem

$$F\colon h \to \frac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{12h}$$

w.r.t. changes in the step size h is

$$\kappa_{\text{abs}} = \frac{|F(h + \delta h) - F(h)|}{|\delta h|}$$

To $\mathcal{O}(\delta h)$ accuracy

$$F(h + \delta h) - F(h) = \frac{-f_2' + 4f_1' + 4f_{-1}' - f_{-2}'}{6h} \delta h \cong f_0' \frac{\delta h}{h},$$

leading to condition number estimate

$$\kappa_{\text{abs}} \cong \left| \frac{f_0'}{h} \right|.$$

The *overall absolute error* in floating point computation of f_0' using (3) will contain *the truncation error* $4h^4/5!\, f_0^{(5)}$ (error due to use of an approximate analytical expression, i.e. 'like *truncating* a series') *and the rounding error* $\kappa_{\text{abs}}\, \epsilon_{\text{mach}}$ (error due to use of approximations of real numbers, i.e. 'like *rounding* a number'),

$$e(h) = Ah^4 + B\epsilon_{\text{mach}}/h, A = \left| \frac{4}{5!} f_0^{(5)} \right|, B = |f_0'|,$$

with an optimal step size h determined by

$$\frac{\mathrm{d}\varepsilon(h)}{\mathrm{d}h} = 0 \Rightarrow h_{\text{opt}} = \left(\frac{15\,\epsilon_{\text{mach}}}{2} \left| \frac{f_0'}{f_0^{(5)}} \right| \right).$$

(Grading note: Formula (3) ~30%, error analysis ~70%)

2. Write an algorithm to evaluate $f(x) = (1 - \cos x)/x$ with minimal loss of precision in floating point arithmetic.

Solution. $f(x)$ numerator can lead to loss of precision when $\cos x \cong 1$, and the denominator is 0 at $x = 0$, but

$$\lim_{x \to 0} f(x) = 0,$$

using $1 - \cos x = 2\sin^2(x/2)$.

Algorithm 1

function $f(x)$:
  if abs$(x) < \epsilon_{\text{mach}}$ then
    return x
  else
    $x2 = 0.5x,\ y = \sin(x2)$
    return $y^2/x2$

3. Let $p \in (0, \infty)$. What is the value of

$$x = \sqrt{p + \sqrt{p + \ldots + \sqrt{p + \ldots}}}\ ?$$

Solution. Define sequence $x_{n+1} = \sqrt{p + x_n}$, $x_0 = 0$, and note that $x = \lim_{n \to \infty} x_n$, and x is fixed point of $g(x) = \sqrt{p + x}$. Solve

$$x = \sqrt{p + x}$$

to find

$$x = \frac{1 + \sqrt{1 + 4p}}{2}.$$

4. Show that $\|\|: \mathbb{R}^{m \times m} \to \mathbb{R}_+$ defined as

$$\|A\| = \sum_{i=1}^{m} \sum_{j=1}^{m} |a_{ij}|,$$

is a matrix norm, and not subordinate to any vector norm.

Solution. Verify norm properties:

a) $\|A\| = 0 \Rightarrow A = 0$: $\sum_{i=1}^m \sum_{j=1}^m |a_{ij}| = 0$ sum of positive quantities, hence $a_{ij} = 0$, i.e. $A = 0$

b) $\|\alpha A\| = |\alpha|\,\|A\|$: $\|\alpha A\| = \sum_{i=1}^m \sum_{j=1}^m |\alpha a_{ij}| = |\alpha| \sum_{i=1}^m \sum_{j=1}^m |a_{ij}| = |\alpha|\,\|A\|$

c) $\|A + B\| \geqslant \|A\| + \|B\|$: $\sum_{i=1}^m \sum_{j=1}^m |a_{ij} + b_{ij}| \geqslant \sum_{i=1}^m \sum_{j=1}^m |a_{ij}| + \sum_{i=1}^m \sum_{j=1}^m |b_{ij}| = \|A\| + \|B\|$

The matrix norm $\| \|$ is subordinate to a vector norm $\| \|_v$ if

$$\|A\| = \sup_{\|x\|_v = 1} \|Ax\|_v.$$

By contradiction: assume $\| \|$ is subordinate to $\| \|_v$, take $A = I$, and let u denote a unit norm vector ($\|u\|_v = 1$) such that $\|I\| = \sup_{\|x\|_v = 1} \|Ix\|_v = 1$, but $\|I\| = \sum_{i=1}^m \sum_{j=1}^m |e_{ij}| = m$, contradiction for $m > 1$.

5. For given $A \in \mathbb{R}^{m \times n}$, $\alpha \in \mathbb{R}_+$, define $F: \mathbb{R}^n \to \mathbb{R}_+$ by

$$F(x) = \|Ax - b\|_2^2 + \alpha \|x\|_2^2.$$

Prove that solving

$$\min_x F(x),$$

is equivalent to solving

$$(A^T A + \alpha I)\, x = A^T b. \tag{4}$$

For x solution of (1) compute $F(x+h)$ $(h \in \mathbb{R}_+)$ in terms of $F(x), \alpha, h, A$.

Solution. Write

$$F(x) = (x^T A^T - b^T)(Ax - b) + \alpha x^T x = x^T A^T A x - b^T A x - x^T A^T b + b^T b + \alpha x^T x,$$

and compute

$$\nabla F(x) = 2 A^T A x - 2 A^T b + 2 \alpha x.$$

$F(x)$ is convex, hence stationary points x satisfying $\nabla F(x) = 0$ are minima so

$$(A^T A + \alpha I)\, x = A^T b \Leftrightarrow \min_x F(x).$$

Compute $F(x+h)$ by Taylor series expansion

$$F(x+h) = F(x) + h^T (A^T A + \alpha I) h.$$

In the above, the linear term in h is null since x is chosen such that $F(x)$, and there are no terms higher than quadratic since F is quadratic in x.

6. Determine the end conditions for a cubic spline interpolation $S(x)$ that minimize

$$\int_a^b [S''(x)]^2 \, \mathrm{d}x.$$

Solution. Let $a = x_0, x_1, \ldots, x_n = b$ denote the knots of the spline interpolation of function $y \in C^2([x_0, x_n])$, $y(x_i) \equiv y_i$, $i = 0, \ldots n$. Recall that a cubic spline is defined as

$$S(x) = S_i(x) = a_i (x - x_{i-1})^3 + b_i (x - x_{i-1})^2 + c_i (x - x_{i-1}) + y_{i-1} \text{ for } x \in [x_{i-1}, x_i],$$

already satisfying interpolation conditions $S(x_{i-1}) = y_{i-1}$, and requiring $3n$ conditions to determine (a_i, b_i, c_i) for $i = 1, \ldots, n$. Imposing $S_i(x_i) = y_i$, $i = 1, \ldots, n$, and $S_i'(x_i) = S_{i+1}'(x_i)$, $S_i''(x_i) = S_{i+1}''(x_i)$ for $i = 1, \ldots, n-1$, gives $3n - 2$ conditions, hence 2 end conditions are arbitrary.

The functional

$$I(S) = \int_a^b [S''(x)]^2 \, \mathrm{d}x \geqslant 0,$$

is a measure of the overall curvature of the spline interpolation, which should be bounded by the curvature of the interpolated function $y(x)$ itself to reduce approximation error away from the interpolation nodes, i.e. the inequality

$$I(S) \leqslant I(y)$$

should hold. Establish this by introducing the error function $E(x) = y(x) - S(x)$, and compute

$$I(y) = \int_a^b [y''(x)]^2 \, \mathrm{d}x = \int_a^b [E''(x) + S''(x)]^2 \, \mathrm{d}x = I(S) + I(E) + 2\int_a^b S''(x)E''(x) \, \mathrm{d}x.$$

Since $I(E) \geqslant 0$, proving that

$$J = \int_a^b S''(x)E''(x) \, \mathrm{d}x = 0,$$

would establish $I(S) \leqslant I(y)$ (note that this is a scalar product of the curvature, and the above imposes orthogonality of the spline curvature to the error curvature). Separate the integral over knot subintervals and integrate by parts ($u = S'' \Rightarrow \mathrm{d}u = S'''\mathrm{d}x$, $\mathrm{d}v = E''\mathrm{d}x \Rightarrow v = E'$)

$$J = \int_a^b S''(x)\, E''(x) \, \mathrm{d}x = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} S_i''(x)\, E_i''(x) \mathrm{d}x = \sum_{i=1}^n \left\{ [S''(x)E'(x)]_{x=x_{i-1}}^{x=x_i} - \int_{x_{i-1}}^{x_i} E'(x)S_i'''(x) \, \mathrm{d}x \right\}$$

Over interval $[x_{i-1}, x_i]$, $S_i''' = 6a_i$, constant, hence

$$\int_{x_{i-1}}^{x_i} E'(x)S_i'''(x) \, \mathrm{d}x = 6a_i \int_{x_{i-1}}^{x_i} E'(x) \, \mathrm{d}x = 6a_i(E(x_i) - E(x_{i-1})) = 0,$$

since there is no error at nodes (interpolation condition), leading to

$$J = \sum_{i=1}^n [S''(x)E'(x)]_{x=x_{i-1}}^{x=x_i}$$

$$\begin{aligned}
J \;=\; & S''(x_n)E'(x_n) - S''(x_{n-1})E'(x_{n-1}) + \\
& S''(x_{n-1})E'(x_{n-1}) - S''(x_{n-2})E'(x_{n-2}) + \\
& \ldots + \\
& S''(x_1)E'(x_1) - S''(x_0)E'(x_0) = S''(x_n)E'(x_n) - S''(x_0)E'(x_0),
\end{aligned}$$

and J can be made null by choosing end conditions $S''(x_0) = 0$, $S''(x_n) = 0$.

$$I =$$