Practice and Instructions for the Final Exam

Information and rules for the final

- The final exam is 5:10 to 7pm on Thursday, December 22.
- No electronic devices may be visible during the exam.
- The exam is "closed book". No materials are allowed except for one "cheat sheet", which is one standard piece of paper that you may prepare in advance with any information you want.
- Please explain all answers with a few words or sentences. You may lose points for a formula or answer with no explanation.
- You will get 25% credit for any question left blank. Points may be deducted for wrong answers, which can bring the credit to zero.
- Please cross out any answers you think are wrong. You may lose points for wrong answers even if you also give the right answer.
- Write answers in an exam book ("blue book").

True/False

In each case, say whether the statement is true or false and give few words or sentences to justify your answer.

1. Consider an IEEE floating point arithmetic format with n fraction bits and k exponent bits. The machine precision ϵ_{mach} of this arithmetic depends on n but not on k.
2. Suppose you are trying to find the minimum of a function $V(x)$ of n variables $(x_1, \cdots, x_n)$. The negative gradient $p = -\nabla V(x)$ is a descent direction.
3. Consider a first order accurate approximation $A(h) = A + O(h)$ and a second order accurate approximation $B(h) = A + O(h^2)$. If $h = .1$ then $B(h)$ is necessarily more accurate than $A(h)$.
4. If A is an $n \times n$ matrix with $\kappa(A) = \|A\| \left\|A^{-1}\right\| = 10^{20}$, then the eigenvalues of A cannot be computed accurately in double precision floating point.
5. If A is an $n \times n$ matrix with $\kappa(A) = \|A\| \left\|A^{-1}\right\| = 10^{20}$, then the solution x of $Ax = b$ cannot be computed accurately for any $b \neq 0$ in double precision floating point arithmetic.

Multiple Choice

In each case, select the correct answer and justify your choice with a few words or sentences.

1. If f_n satisfy the recurrence relation $f_{n+2} = f_n + f_{n+1}$, and f_0 and f_1 are given. Which of the following is true if calculations are done in double precision floating point? Select all that apply.
   (a) The result is computed to 1% relative accuracy if $|f_n| < 10^{10}$.
   (b) The result is computed to 1% relative accuracy if $|f_n| < 10^{10}$ and $f_0 > 0$ and $f_1 > 0$.
   (c) $f_n \approx Cr^n$, where $r = \frac{1+\sqrt{5}}{2}$ satisfies the equation $r^2 = r + 1$ and f_0 and f_1 are arbitrary.

(d) The result is computed to 1% relative accuracy if $|f_n| < 10^{10}$ and $f_0 > 0$ and $f_1 > 0$ and $f_0 > 1$.

2. Which of the following computes the matrix product $C = AB$ fastest when $n = 1000$? Suppose A and B are $n \times n$ matrices.

(a)
```
C = np.zeros([n,n])
for i in range(n):
   for j in range(n):
      for k in range(n):
         C[i,j] += A[i,k]*B[k,j]
```

(b)
```
C = np.zeros([n,n])
for i in range(n):
   for j in range(n):
      C[i,j] = np.dot( A[i,:], B[:,j])
```

(c)
```
C = A @ B
```

(d)
```
C = A * B
```

3. Consider the approximation

$$e^x \approx A(x) = \sum_{n=0}^{1000} \frac{x^n}{n!} .$$

Which of the following is not true about this approximation? Select all that apply, as there may be more than one. You may use the fact that $e^{50} = 5.18 \cdots \times 10^{21}$.

(a) $A(50)$ approximates e^{50} to high relative accuracy when computed using double precision floating point arithmetic.

(b) $A(-50)$ approximates e^{-50} to high relative accuracy, if the operations are performed exactly, without any rounding error.

(c) $A(-50)$ approximates e^{-50} to high relative accuracy when computed using double precision floating point arithmetic.

(d) When $\widehat{A}(50) \approx A(50)$ is computed in double precision floating point and e^{50} is computed exactly, then $\widehat{A}(50$ has reasonable absolute accuracy in the sense that $\left|\widehat{A}(50) - e^{50}\right| < .01$.

(e) When $A(50)$ and e^{50} are computed exactly $A(50)$ has reasonable absolute accuracy in the sense that $\left|A(50) - e^{50}\right| < .01$.

Full answer

1. Suppose a code produces estimates $A(h)$ from an approximation that has order of accuracy p. Suppose the code produces the following output:

h	A(h)
.4	24.07
.2	23.65
.1	23.44

Use asymptotic error analysis to estimate the order of accuracy and the answer more accurately (hopefully) than any of the numbers in the table.

2. Suppose $n \times n$ matrices A and M are known with $AM = I$. Suppose C is a matrix that is close to A and we want to solve the equation $Cx = b$ (find x given b). Use perturbation theory to find an approximation to x that is more accurate (hopefully) than Mb. The approximation should involve multiplying matrices but not matrix inversion.

3. Consider the linear least squares approximation problem

$$\min_x \|Ax - b\|_2 \ .$$

Find a formulation of this as a quadratic minimization problem

$$\min_x F(x) \ , \quad F(x) = x^T M x - c^T x \ .$$

Describe a gradient descent algorithm for the quadratic minimization problem that is convergent if the step size/learning rate is small enough. Find a formula for $\nabla F(x)$.

4. Consider the norm constrained least squares problem

$$\min_{\|x\|_2 \leq R} \|Ax - b\|_2 \ .$$

Find an algorithm to solve this using the SVD of A. Note: You need to check whether the constraint is binding for the unconstrained minimum.

5. Suppose we want to estimate $f(h)$ using $f(0)$, $f(-h)$ and $f'(0)$. Suppose $f(x)$ is a smooth function of x. Find an approximation of the form $f(h) \approx af(-h) + bf(0) + cf'(0)$ with the highest possible order of accuracy. The coefficients a, b, and c may depend on h but not on f.

6. Consider the time stepping method for solving the ODE system $\dot{x} = f(x)$ with time step Δt and $x_k \approx x(t_k)$, $t_k = k\Delta t$

$$y_{k+1} = x_k + \Delta t f(x_k) \ , \quad x_{k+1} = x_k + \Delta t \tfrac{1}{2} \left[f(x_k) + f(y_{k+1}) \right] \ .$$

Estimate the local truncation error and find the overall accuracy of this method.