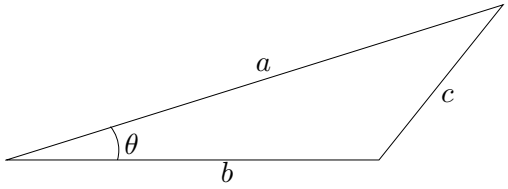
L. Vandenberghe 12/10/08
EE103

Final Exam Solutions

Problem 1 (10 points).



The length of one side of a triangle (c) can be calculated from the lengths of the two other sides (a and b) and the opposing angle (θ) by the formula

$$c = \sqrt{a^2 + b^2 - 2ab\cos\theta}. \tag{1}$$

Two equivalent expressions are

$$c = \sqrt{(a-b)^2 + 4ab\,(\sin(\theta/2))^2} \tag{2}$$

and

$$c = \sqrt{(a+b)^2 - 4ab\,(\cos(\theta/2))^2}. \tag{3}$$

(The equivalence of the three formulas follows from the identities $\cos\theta = 1 - 2\,(\sin(\theta/2))^2$ and $\cos\theta = -1 + 2\,(\cos(\theta/2))^2$.)
Which of the three formulas gives the most stable method for computing c if $a \approx b$ and θ is small? For simplicity you can assume that the calculations are exact, except for a small error in the evaluation of the cosine and sine functions. Explain your answer.

Solution. Expressions (1) and (3) suffer from cancellation; expression (2) does not.
Cancellation occurs when two numbers are subtracted that are almost equal, and one or both are subject to error. Therefore cancellation occurs in the subtraction in (1) and (3). In (2) we also subtract two almost equal numbers a and b, but they are not subject to error (under the assumptions of the problem).

Problem 2 (10 points) How many IEEE double precision floating-point numbers are contained in the following intervals?

1. The interval $[1/2, 3/2)$.
2. The interval $[3/2, 5/2)$.

Explain your answer.

Solution.

1. $3 \cdot 2^{51}$. The floating-point representations of the numbers 1/2, 1, and 3/2 are

$$1/2 = (.100\cdots0)_2 \cdot 2^0, \qquad 1 = (.100\cdots0)_2 \cdot 2^1, \qquad 3/2 = (.110\cdots0)_2 \cdot 2^1.$$

   There are 2^{52} floating-point numbers in $[1/2, 1)$ and 2^{51} in $[1, 3/2)$.

2. $3 \cdot 2^{50}$. The floating-point representations of the numbers 3/2, 2, and 5/2 are

$$3/2 = (.1100\cdots0)_2 \cdot 2^1, \qquad 2 = (.1000\cdots0)_2 \cdot 2^2, \qquad 5/2 = (.1010\cdots0)_2 \cdot 2^2.$$

   There are 2^{51} floating-point numbers in $[3/2, 2)$ and 2^{50} floating-point numbers in $[2, 5/2)$.

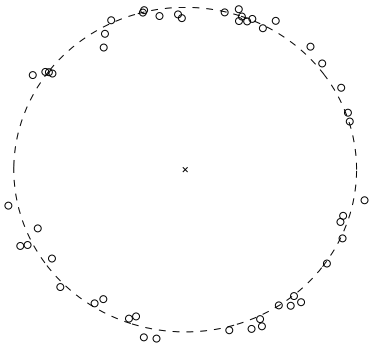
Problem 3 (10 points). Explain how you would solve the following problem using the Gauss-Newton algorithm. Fit a circle

$$(u - u_c)^2 + (v - v_c)^2 = R^2$$

to m given points (u_i, v_i) in a plane. In other words, determine u_c, v_c, R such that

$$(u_i - u_c)^2 + (v_i - v_c)^2 \approx R^2, \quad i = 1, \ldots, m.$$

The variables are the coordinates of the center u_c, v_c, and the radius R.



Your description should include:

- The cost function that you minimize.
- The matrix A and the vector b in the least-squares problem

$$\text{minimize} \quad \|Ax - b\|$$

  that you solve at each iteration.

You do not have to discuss the choice of starting point, the stopping criterion, and the line search.

Solution.

- We use as variables $x = (R, u_c, v_c)$, and minimize the function

$$g(x) = \sum_{i=1}^{m} r_i(x)^2, \qquad r_i(x) = R^2 - (u_i - u_c)^2 - (v_i - v_c)^2.$$

- At each iteration we minimize $\|Ax - b\|$ where

$$A = \begin{bmatrix} \nabla r_1(\hat{x})^T \\ \vdots \\ \nabla r_m(\hat{x})^T \end{bmatrix}, \qquad b = A\hat{x} - \begin{bmatrix} r_1(\hat{x}) \\ \vdots \\ r_m(\hat{x}) \end{bmatrix}$$

  and $\hat{x} = (\hat{R}, \hat{u}_c, \hat{v}_c)$ is the current iterate. The gradient of r_i is

$$\nabla r_i(\hat{x}) = 2 \begin{bmatrix} \hat{R} \\ u_i - \hat{u}_c \\ v_i - \hat{v}_c \end{bmatrix}$$

  and therefore

$$A = 2 \begin{bmatrix} \hat{R}\mathbf{1} & u - \hat{u}_c\mathbf{1} & u - \hat{v}_c\mathbf{1} \end{bmatrix}$$

  where $\mathbf{1}$ is the m-vector with all its entries equal to one, and u and v are the m-vectors with entries u_i and v_i.

Problem 4 (10 points). If A is an $m \times n$-matrix with a zero nullspace, and D is an $m \times m$ diagonal matrix with positive diagonal elements, then the coefficient matrix of the equation

$$\begin{bmatrix} D^2 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix}$$

is nonsingular. Therefore the equation has a unique solution $\hat{x}$, $\hat{y}$.

1. Show that $\hat{x}$ is the solution of the optimization problem

$$\begin{array}{ll} \text{minimize} & \|Dx - D^{-1}b\|^2 \\ \text{subject to} & A^T x = c. \end{array}$$

2. Show that $\hat{y}$ is the solution of the optimization problem

$$\text{minimize} \quad \|D^{-1}(Ay - b)\|^2 + 2c^T y.$$

   (Hint: set the gradient of the cost function to zero.)

3. Describe an efficient method, based on the QR factorization of $D^{-1}A$, for computing $\hat{x}$ and $\hat{y}$. Clearly state the different steps in your algorithm, the complexity of each step (number of flops for large m, n), and the total complexity.

Solution. We first derive expressions for $\hat{x}$ and $\hat{y}$. From the first equation, $\hat{x} = D^{-2}(b - A\hat{y})$. Substituting this in the second equation gives $A^T D^{-2} b - A^T D^{-2} A\hat{y} = c$. Hence

$$\begin{aligned}
\hat{y} &= (A^T D^{-2} A)^{-1}(A^T D^{-2} b - c) \\
\hat{x} &= D^{-2}(b - A\hat{y}) \\
&= D^{-2}\left(b - A(A^T D^{-2} A)^{-1}(A^T D^{-2} b - c)\right).
\end{aligned}$$

1. This problem can be reduced to a least-norm problem via a change of variables $z = Dx - D^{-1}b$, $x = D^{-1}(z + D^{-1}b)$:

$$\begin{array}{ll} \text{minimize} & \|z\|^2 \\ \text{subject to} & A^T D^{-1} z = c - A^T D^{-2} b. \end{array}$$

   From the theory of least-norm problems we know that the solution is

$$z = D^{-1}A(A^T D^{-2} A)^{-1}(c - A^T D^{-2} b).$$

   Therefore

$$x = D^{-1}(z + D^{-1}b) = D^{-2}\left(A(A^T D^{-2} A)^{-1}(c - A^T D^{-2} b) + b\right),$$

   the same solution as $\hat{x}$ derived above.

2. We write the cost function as

$$\begin{aligned}
f(y) &= (D^{-1}(Ay - b))^T (D^{-1}(Ay - b)) + 2c^T y \\
&= y^T A^T D^{-2} A y - 2b^T D^{-2} A y + b^T D^{-2} b + 2c^T y.
\end{aligned}$$

   This is has the general form of a quadratic function $g(y) = y^T P y + q^T y + r$ with gradient $\nabla g(y) = 2Py + q$. Setting the gradient of f with respect to y equal to zero therefore gives

$$\nabla f(y) = 2A^T D^{-2} A y - 2A^T D^{-2} b + 2c = 0,$$

   and

$$y = (A^T D^{-2} A)^{-1}(A^T D^{-2} b - c),$$

   the same solution as $\hat{y}$ derived above.

3. We use the QR factorization $D^{-1}A = QR$. The expressions for $\hat{x}$ and $\hat{y}$ then reduce to

$$\begin{aligned}
\hat{y} &= (R^T R)^{-1}(R^T Q^T D^{-1} b - c) \\
&= R^{-1}(Q^T D^{-1} b - R^{-T} c) \\
\hat{x} &= D^{-2}(b - A\hat{y}) \\
&= D^{-2}(b - QR\hat{y})
\end{aligned}$$

   The different steps are

   - Compute $D^{-1}A$ (mn flops) and factor it as $D^{-1}A = QR$ ($2mn^2$ flops).
   - Compute $D^{-1}b$ (m flops) and $u = Q^T D^{-1} b$ ($2mn$ flops).
   - Solve $R^T v = c$ by forward substitution (n^2 flops) and compute $w = u - v$ (n flops).
   - Solve $R\hat{y} = w$ by backsubstitution (n^2 flops).
   - Compute Qw ($2mn$ flops) and $\hat{x} = D^{-2}(b - Qw)$ ($3m$ flops.)

   The total for large m, n is $2mn^2$.

Problem 5 (10 points). An $m \times n$-matrix A is given in factored form

$$A = UDV^T$$

where U is $m \times n$ and orthogonal, D is $n \times n$ and diagonal with nonzero diagonal elements, and V is $n \times n$ and orthogonal. Describe an efficient method for solving the least-squares problem

$$\text{minimize} \quad \|Ax - b\|^2.$$

'Efficient' here means that the complexity is substantially less than the complexity of the standard method based on the QR factorization. What is the cost of your algorithm (number of flops for large m and n)? (Note: we assume that U, D, V are given; you are not asked to include the cost of computing these matrices in the complexity analysis.)

Solution.
We use the following properties:

- U is orthogonal. Hence by definition $U^TU = I$.
- D is diagonal with nonzero diagonal elements. Therefore it is nonsingular and its inverse D^{-1} is the diagonal matrix with $1/D_{ii}$ on the diagonal.
- V is orthogonal and square. Therefore V is nonsingular and $V^{-1} = V^T$.

The solution of the least-squares problem satisfies the normal equations

$$A^TAx = A^Tb.$$

Replacing A with UDV^T and using the property $U^TU = I$ gives

$$VD^2V^Tx = VDU^Tb.$$

Multiplying with $(VD^2V^T)^{-1} = VD^{-2}V^T$ on the left gives

$$x = VD^{-1}U^Tb.$$

The cost of evaluating $VD^{-1}U^Tb$ is $2mn + n + 2n^2 \approx 2mn + 2n^2$ flops. ($2mn$ for the matrix-vector product U^Tb, n for the multiplication with D^{-1}, and $2n^2$ for the matrix-vector product with V.

Problem 6 (10 points) Let L be a nonsingular $n \times n$ lower triangular matrix with elements L_{ij}. Show that

$$\kappa(L) \geq \frac{\max_{i=1,\ldots,n} |L_{ii}|}{\min_{j=1,\ldots,n} |L_{jj}|}.$$

Solution. By definition $\kappa(L) = \|L\|\|L^{-1}\|$. To bound κ we use the inequalities

$$\|L\| \geq \frac{\|Lx\|}{\|x\|}, \qquad \|L^{-1}\| \geq \frac{\|L^{-1}y\|}{\|y\|},$$

where x and y are nonzero vectors.
We choose $x = e_i$ (the ith unit vector, *i.e.*, a vector of zeros except for an element equal to one in position i). Then Lx is the ith column of L, and $(Lx)_i = L_{ii}$. Therefore $\|Lx\| \geq |L_{ii}|$. This proves that $\|L\| \geq |L_{ii}|$.
We choose $y = e_j$. If we determine $L^{-1}y$ by forward substitution, we find that $(L^{-1}y)_j = 1/L_{jj}$. Therefore $\|L^{-1}y\| \geq 1/|L_{jj}|$. This proves $\|L^{-1}\| \geq 1/|L_{jj}|$.
Multiplying the two bounds gives $\kappa(L) \geq |L_{ii}|/|L_{jj}|$ for all i and j.

Problem 7 (10 points) Suppose A is an $n \times n$ positive definite matrix. For what values of the scalar β is the matrix

$$\begin{bmatrix} A & -A \\ -A & \beta A \end{bmatrix}$$

positive definite? Explain your answer.

Solution. A matrix is positive definite if and only if it has a Cholesky factorization. Let $A = LL^T$ be the Cholesky factorization of A. We determine the Cholesky factorization of the block matrix:

$$\begin{bmatrix} LL^T & -LL^T \\ -LL^T & \beta LL^T \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix},$$

with L_{11} and L_{22} lower triangular. From the 1,1 block we see that $LL^T = L_{11}L_{11}^T$. Therefore $L_{11} = L$. From the 2,1 block $-LL^T = L_{21}L^T$. Therefore $L_{21} = -L$. Finally, from the 2,2 block

$$(\beta - 1)LL^T = L_{22}L_{22}^T.$$

If $\beta \leq 1$ the matrix on the left is not positive definite, so it cannot be factored as $L_{22}L_{22}^T$. We therefore obtain the condition $\beta > 1$ and $L_{22} = \sqrt{\beta - 1}L$.
An alternative approach is to use the definition of positive definite matrix. The block matrix is positive definite if and only if

$$\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} A & -A \\ -A & \beta A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} > 0$$

for all nonzero (x, y). We have

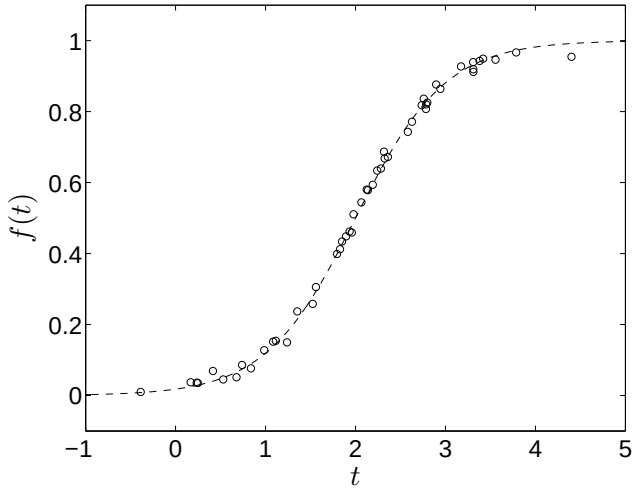
$$\begin{aligned}\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} A & -A \\ -A & \beta A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= x^T A x - 2y^T A x + \beta y^T A y \\ &= (x-y)^T A (x-y) + (\beta - 1) y^T A y.\end{aligned}$$

This is positive for all nonzero x, y if and only if $\beta > 1$.

Problem 8 (10 points). The figure shows $m = 50$ points (t_i, y_i) as circles. These points are well approximated by a function of the form

$$f(t) = \frac{e^{\alpha t + \beta}}{1 + e^{\alpha t + \beta}}.$$

(An example is shown in dashed line).



Formulate the following problem as a linear least-squares problem. Find values of the parameters α, β such that

$$\frac{e^{\alpha t_i + \beta}}{1 + e^{\alpha t_i + \beta}} \approx y_i, \qquad i = 1, \ldots, m, \tag{4}$$

You can assume that $0 < y_i < 1$ for $i = 1, \ldots, m$.
Clearly state the error function you choose to measure the quality of the fit in (4), and the matrix A and the vector b of the least-squares problem.

Solution. The inverse of the nonlinear function $h(x) = e^x/(1+e^x)$ is $h^{-1}(y) = \log(y/(1-y))$, *i.e.*,

$$y = \frac{e^x}{1+e^x} \qquad \Longleftrightarrow \qquad x = \log\left(\frac{y}{1-y}\right).$$

Applying this nonlinear transformation to the two sides of (4) gives a linear set of equations

$$\alpha t_i + \beta \approx \log\left(\frac{y_i}{1-yi}\right), \qquad i = 1, \ldots, m.$$

This means that if we use as error function

$$\sum_{i=1}^{m} \left(\alpha t_i + \beta - \log(\frac{y_i}{1-y_i})\right)^2$$

we get a linear least-squares problem with

$$x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \qquad A = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} \log(y_1/(1-y_1)) \\ \log(y_2/(1-y_2)) \\ \vdots \\ \log(y_m/(1-y_m)) \end{bmatrix}.$$