WASP MATHEMATICS FOR MACHINE LEARNING 2024
Instructor: Anders Forsgren, KTH

Homework assignment 1
Due Tuesday March 5 2024

There will be two homework assignments, out of which this is the first one. Each homework assignment will have two subsections with questions corresponding to 30 points each, in this case "Linear algebra" and "Calculus". Each student is expected to do his/her best, given his/her level. To pass the course, a sufficient overall performance is required. A passing grade is guaranteed if at least 10 points on each subsection is obtained, giving a total of at least 40 points.

General rules for the homework assignments:

- Each student has to hand in individual solutions.
- You may make use of the textbook and other literature. If you make use of literature, give references in your solutions.
- You may cooperate. Discussions are encouraged. If you receive help from someone, say so in your solutions.
- You must hand in your own original solutions. You are *not* allowed to make use of solutions made by others in any form.
- Hand in solutions to an assignment as one pdf file via Canvas.

The textbook should be sufficient as a reference for almost all problems. You may want to consult other literature for a few problems.

For linear algebra, you might find Lloyd N. Trefethen, David Bau III, *Numerical linear algebra* helpful.

For calculus, it would depend on your background what literature you might need in addition to the textbook.

Linear algebra (30 points)

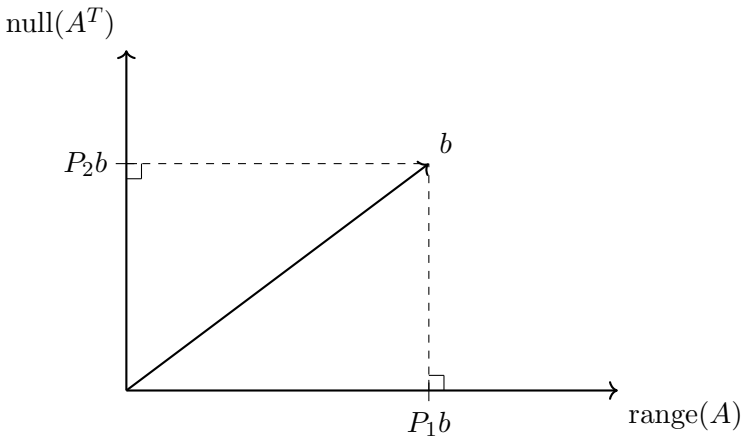
1. (2 points) Find a matrix A, whose eigenvalues are $\lambda_1 = -3$ and $\lambda_2 = 1$, and for which the corresponding eigenvectors are $(1\ {-1})^T$ and $({-2}\ \ 1)^T$ respectively.

2. (2 points) Let $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. What are the eigenvalues of A? What are the eigenvalues of A^2? What is the geometric meaning of the linear transformations A, A^2, A^3 and A^4?

3. (3 points) Let $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = n$, and let $b \in \mathbb{R}^m$. Then,

$$b = A(A^TA)^{-1}A^Tb + (I - A(A^TA)^{-1}A^T)b = P_1b + P_2b,$$

for $P_1 = A(A^TA)^{-1}A^T$ and $P_2 = I - A(A^TA)^{-1}A^T$. Show that P_1 gives the orthogonal projection onto range(A) and that P_2 gives the orthogonal projection onto null(A^T). Use these results to conclude that $\hat{x} = (A^TA)^{-1}A^Tb$ is the optimal solution to $\min \|Ax - b\|_2^2$.



4. (3 points) Let $A \in \mathbb{R}^{n\times n}$ with $\text{rank}(A) = n$. Assume that an LU-decomposition of A is given, i.e., $PAQ = LU$ for $L \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{n\times n}$, with L unit lower triangular, i.e., triangular with ones on the diagonal, U upper triangular, and P and Q are permutation matrices. Show that $\det(A) = \prod_{i=1}^{n} u_{ii}$. Explain why an LU-decomposition would be helpful for solving a system of linear equations $Ax = b$.

Note (which is relevant but not needed to solve the problem): A step in an LU-decomposition is analogous to Gaussian elimination. In the first step, we may write A in partitioned form as

$$A = \begin{pmatrix} a & b^T \\ c & D \end{pmatrix},$$

where a is the (1,1)-element. We assume $a \neq 0$, otherwise row and column permutations may give $a \neq 0$ unless $A = 0$, in which case the algorithm terminates with $L = I$ and $U = A = 0$. For $a \neq 0$,

$$\begin{aligned} A = \begin{pmatrix} a & b^T \\ c & D \end{pmatrix} &= \begin{pmatrix} 1 \\ \frac{1}{a}c \end{pmatrix} \begin{pmatrix} a & b^T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & D - \frac{1}{a}cb^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \frac{1}{a}c & I \end{pmatrix} \begin{pmatrix} a & b^T \\ 0 & D - \frac{1}{a}cb^T \end{pmatrix}. \end{aligned}$$

This gives one column of L as $(1 \ \frac{1}{a}c^T)^T$, one row of U as $(a \ b^T)$, plus a matrix $D - (1/a)cb^T$, which has one row and column less than A. We may now apply the same procedure on $D - (1/a)cb^T$. Repeating this procedure r steps will give L and U, where r is the rank of A.

5. (3 points) Let A be given by

$$A = \begin{pmatrix} 1 & 2 & 3 & 1 & 1 \\ -1 & -1 & -1 & -3 & -2 \\ 1 & 3 & 5 & -1 & 0 \end{pmatrix}.$$

What is the rank of A? Find a matrix whose columns form a basis for range(A^T). Find a matrix whose columns form a basis for null(A).

It may be helpful to make use of an LU-decomposition of A, where $A = LU$ for

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} 1 & 2 & 3 & 1 & 1 \\ 0 & 1 & 2 & -2 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

6. (3 points) Let A be as in the previous question, and let $b = (1\ 0\ 2)^T$.

Find all solutions to $Ax = b$. This means that your task is to find a particular solution $\bar{x}$ and a matrix Z whose columns form a basis for the nullspace of A. We can then obtain all solutions as $x = \bar{x} + Zv$, where v is an arbitrary vector of the appropriate dimension.

Make use of your solution from the previous question.

7. (3 points) For a matrix $A \in \mathbb{R}^{m\times n}$ of rank r, the *singular value decomposition* of A, SVD, is given by $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times m}$ is orthogonal, i.e., $U^TU = UU^T = I_m$, $V \in \mathbb{R}^{n\times n}$ is orthogonal, i.e., $V^TV = VV^T = I_n$, and $\Sigma \in \mathbb{R}^{m\times n}$ has elements $\sigma_{11} \geq \sigma_{22} \geq \ldots \geq \sigma_{rr} > \sigma_{r+1,r+1} = \ldots \sigma_{\min\{m,n\},\min\{m,n\}} = 0$, $\sigma_{ij} = 0$, $i \neq j$.

Suppose that the matrix $A \in \mathbb{R}^{n\times n}$ has an SVD given by $A = U\Sigma V^T$. Let the $2n \times 2n$ matrix M be given by

$$M = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}.$$

Find an eigenvalue decomposition (the one in the spectral theorem) of M.

Note (which is relevant but not needed to solve the problem): It is recommended to read more about the singular value decomposition and its use in image compression in Sections 4.5 and 4.6 of the textbook.

8. (3 points) Let H and c be given by

$$H = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 4 & 2 \\ -1 & 2 & 6 \end{pmatrix}, \quad \text{and} \quad c = \begin{pmatrix} 0 \\ 2 \\ -3 \end{pmatrix}.$$

This H is symmetric and positive definite. A *Cholesky decomposition* of H has been computed, so that $H = R^TR$, for

$$R = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Your task is to use the result of the decomposition to find x such that $Hx + c = 0$. First show that if $Hx + c = 0$ and $H = R^TR$, then $Rx = y$ for $R^Ty = -c$. Then use this result to find x. Solving by R^T and R is straightforward, because R is upper triangular.

Note (which is relevant but not needed to solve the problem): Cholesky decomposition is equivalent to an LDL^T- decomposition, which may be viewed as a special

case of LU-decomposition for the case when the matrix H is symmetric and positive definite. The LDL^T-decomposition gives a unit lower triangular matrix L and diagonal matrix D, with positive diagonal elements, such that $H = LDL^T$. Analogous to the description of LU-decomposition of question 4, we may write H in partitioned form as

$$H = \begin{pmatrix} a & b^T \\ b & C \end{pmatrix},$$

where a is the (1,1)-element. We must have $a > 0$, otherwise H would not be positive definite. (We have $a = e_1^T H e_1$, so positive definiteness gives $e_1^T H e_1 = a > 0$.) Then,

$$\begin{aligned} H &= \begin{pmatrix} a & b^T \\ b & C \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{a}b \end{pmatrix} a \begin{pmatrix} 1 & \frac{1}{a}b^T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & C - \frac{1}{a}bb^T \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ \frac{1}{a}b & I \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & C - \frac{1}{a}bb^T \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{a}b^T \\ 0 & I \end{pmatrix}. \end{aligned}$$

This gives one column of L as $(1 \ \frac{1}{a}b^T)^T$, one diagonal element of D as a, plus a matrix $C - (1/a)bb^T$, which is a symmetric matrix with one row and column less than H. We may now apply the same procedure on $C - (1/a)bb^T$. The positive definiteness of H implies $a > 0$ and $C - (1/a)bb^T$ positive definite, so that the procedure may be repeated for $C - (1/a)bb^T$. Repeating this procedure n steps will give L and D such that $H = LDL^T$. As $d_{ii} > 0$, $i = 1, \ldots, n$, we may define $D^{1/2} = \mathrm{diag}(\sqrt{d_{ii}})$. Then,

$$H = LDL^T = LD^{1/2}D^{1/2}L^T = (D^{1/2}L^T)^T D^{1/2}L^T = R^T R,$$

for $R = D^{1/2}L^T$. Thus, $H = R^T R$, for R upper triangular with strictly positive diagonal elements.

9. (2 points) Assume that $A \in \mathbb{R}^{n\times n}$ is skew-symmetric, which means that $A^T = -A$. What is trace(A)? In addition, let $x \in \mathbb{R}^n$. What is $x^T A x$?

10. (2 points) Assume that $A \in \mathbb{R}^{3\times 3}$ has eigenvalues 1, –2, –3. What is trace(A^2)? What is det(A^{-T})? (Here A^{-T} denotes the transpose of the inverse of A, or equivalently the inverse of the transpose of A.)

11. (2 points) Assume that $A \in \mathbb{R}^{m\times p}$, $B \in \mathbb{R}^{p\times n}$, and $C \in \mathbb{R}^{p\times n}$. Show that if rank$(A) = p$, then $AB = AC$ implies $B = C$. Does the result hold also without the assumption rank$(A) = p$?

12. (2 points) Let $A \in \mathbb{R}^{n\times n}$ and let $x \in \mathbb{R}^n$. Prove that if A is symmetric and positive definite, then trace$(Axx^T) > 0$ for all $x \neq 0$.

Calculus (30 points)

13. (2 points) Compute the gradient and the Hessian of $f(x) = \ln(\sum_{i=1}^n e^{x_i})$.

14. (1 point) Let $f(x) = \frac{1}{1+e^{-x}}$. Prove the identity $f'(x) = f(x)(1 - f(x))$.

15. (2 points) Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x) = \sum_{i=1}^m \ln(1 + e^{a_i^T x})$, where a_i are given vectors. Compute $\nabla f(x)$.

16. (2 points) Here are some functions that are used in neural networks as activation functions: $\sigma(x) = \frac{1}{1+e^{-x}}$, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, $\mathrm{ReLU}(x) = \max\{x, 0\}$. Find the derivatives of each of them. If you have some concerns, write them down.

17. (2 points) For $x \in \mathbb{R}$, consider $f(x) = (\max\{x, 0\})^2$. Find the derivative of f and the second derivative of f. If you haves some concerns, write them down.

18. (1 point) Prove that the function $f(x) = x_1 \ln(\frac{x_2}{x_1})$ satisfies $\langle \nabla f(x), x \rangle = f(x)$.

19. (3 points) Let $f(x) = \frac{1}{2}(x_1+1)^2 + \frac{1}{2}(x_2+1) - \ln(x_1 x_2)$. Find an $x \in \mathbb{R}^2$, with $x_1 > 0$ and $x_2 > 0$, such that $\nabla f(x) = 0$.

Note (which is relevant but not needed to solve the problem): In the optimization lecture, you will learn that the x that you compute is the global minimizer to the optimization problem

$$\underset{x_1>0, x_2>0}{\text{minimize}} \quad \tfrac{1}{2}(x_1+1)^2 + \tfrac{1}{2}(x_2+1) - \ln(x_1 x_2).$$

20. (3 points) Consider the quadratic function $f(x) = \frac{1}{2}x^T H x + c^T x$, where H is a positive definite symmetric $n \times n$ matrix. Show that there is a unique point x^* such that $\nabla f(x^*) = 0$. Further show that f may be written

$$f(x) = \tfrac{1}{2}x^T H x + c^T x = \tfrac{1}{2}(x - x^*)^T H (x - x^*) - \tfrac{1}{2}(x^*)^T H x^*.$$

Finally, use this result to conclude that x^* minimizes f over $\mathbb{R}^n$.

21. (2 points) Consider the quadratic function $f(x) = \frac{1}{2}x^T H x + c^T x$, where H is a symmetric $n \times n$ matrix. Is it possible to choose H and c so that $\nabla f(x) \neq 0$ for all x in $\mathbb{R}^n$?

22. (2 points) Let $f(x_1, x_2) = e^{x_1 x_2^2}$. What is the linear approximation of f at the point (1, 1)? Make a numerical comparison of the values of f and its linear approximation at a couple of points near (1, 1).

23. (2 points) Consider f from the previous question. What is the quadratic approximation of f at this point? Make a numerical comparison of the values of f and its quadratic approximation at the same points that you chose in the previous question. Comment on the results in comparison to the results of the previous question.

24. (2 points) Let $f(x_1, x_2) = \sin(x_1^2 + x_2) + x_2$. Find all critical points of f (those points where the gradient vanishes $\nabla f(x_1, x_2) = 0$).

25. (3 points) Given a function $f : \mathbb{R}^n \to \mathbb{R}$, the level set (in $\mathbb{R}^2$ it is also called the level curve and in $\mathbb{R}^3$ the level surface) is the set $\{x : f(x) = c\}$, where $c \in \mathbb{R}$. We looked at level sets in class when considering a quadratic functions. Give a real world example of the level set. (Hint: Think about geography.)

An important property of the gradient is that the gradient of f at a point $\bar{x}$ is orthogonal to the level set $\{x : f(x) = f(\bar{x})\}$ (that is the level set containing $\bar{x}$).

For a given $\bar{x}$, where $\nabla f(\bar{x}) \neq 0$, let $T(\bar{x})$, the tangent plane to the surface $\{x : f(x) = f(\bar{x})\}$ at $\bar{x}$, be defined by

$$T(\bar{x}) = \{x \in \mathbb{R}^n : \nabla f(\bar{x})^T (x - \bar{x}) = 0\}.$$

Now the problem: Compute the tangent plane associated with $f(x) = 2x_1^2 x_2 + x_3^2$ at the point $\bar{x} = (1\ 1\ 1)^T$.

26. (3 points) Let $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be defined by $f(X) = (\sin(X))^{-1}$, where $\sin(X)$ means elementwise operations, i.e., $(\sin(X))_{ij} = \sin(x_{ij})$. Prove that

$$df_X(H) = -(\sin(X))^{-1}(\cos(X). \cdot H)(\sin(X))^{-1},$$

where "$.\cdot$" denotes componentwise multiplication and

$$df_X(H) = \lim_{t\to 0} \frac{f(X+tH) - f(X)}{t} = \frac{d}{dt} f(X+tH)|_{t=0}.$$

It may be helpful to note that $(X+tH)^{-1}(X+tH) = I$ for all t close to 0, so that

$$\frac{d}{dt}((X+tH)^{-1}(X+tH)) = 0,$$

which by the chain rule gives

$$(\frac{d}{dt}(X+tH)^{-1})(X+tH) + (X+tH)^{-1}H = 0.$$

In particular, for $t = 0$,

$$(\frac{d}{dt}(X+tH)^{-1}|_{t=0})X + X^{-1}H = 0,$$

so that

$$\frac{d}{dt}(X+tH)^{-1}|_{t=0} = -X^{-1}HX^{-1}.$$

Good luck!