

Instructor's Solution Manual for "Linear Algebra and Optimization for Machine Learning"

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY

March 21, 2021

Contents

1	Linear Algebra and Optimization: An Introduction	1
2	Linear Transformations and Linear Systems	17
3	Diagonalizable Matrices and Eigenvectors	35
4	Optimization Basics: A Machine Learning View	47
5	Optimization Challenges and Advanced Solutions	57
6	Lagrangian Relaxation and Duality	63
7	Singular Value Decomposition	71
8	Matrix Factorization	81
9	The Linear Algebra of Similarity	89
10	The Linear Algebra of Graphs	95
11	Optimization in Computational Graphs	101

Chapter 1

Linear Algebra and Optimization: An Introduction

1. For any two vectors \bar{x} and \bar{y} , which are each of length a , show that (i) $\bar{x} - \bar{y}$ is orthogonal to $\bar{x} + \bar{y}$, and (ii) the dot product of $\bar{x} - 3\bar{y}$ and $\bar{x} + 3\bar{y}$ is negative.

(i) The first is simply $\bar{x} \cdot \bar{x} - \bar{y} \cdot \bar{y}$ using the distributive property of matrix multiplication. The dot product of a vector with itself is its squared length. Since both vectors are of the same length, it follows that the result is 0. (ii) In the second case, one can use a similar argument to show that the result is $a^2 - 9a^2$, which is negative.

2. Consider a situation in which you have three matrices A , B , and C , of sizes 10×2 , 2×10 , and 10×10 , respectively.

(a) Suppose you had to compute the matrix product ABC . From an efficiency perspective, would it computationally make more sense to compute $(AB)C$ or would it make more sense to compute $A(BC)$?

(b) If you had to compute the matrix product CAB , would it make more sense to compute $(CA)B$ or $C(AB)$?

The main point is to keep the size of the intermediate matrix as small as possible in order to reduce both computational and space requirements. In the case of ABC , it makes sense to compute BC first. In the case of CAB it makes sense to compute CA first. This type of associativity property is used frequently in machine learning in order to reduce computational requirements.

3. Show that if a matrix A satisfies $A = -A^T$, then all the diagonal elements of the matrix are 0.

Note that $A + A^T = 0$. However, this matrix also contains twice the diagonal elements of A on its diagonal. Therefore, the diagonal elements of A must be 0.

4. Show that if we have a matrix satisfying $A = -A^T$, then for any column vector \bar{x} , we have $\bar{x}^T A \bar{x} = 0$.

Note that the transpose of the scalar $\bar{x}^T A \bar{x}$ remains unchanged. Therefore, we have $\bar{x}^T A \bar{x} = (\bar{x}^T A \bar{x})^T = \bar{x}^T A^T \bar{x} = -\bar{x}^T A \bar{x}$. Therefore, we have $2\bar{x}^T A \bar{x} = 0$.

5. Show that if we have a matrix A , which can be written as $A = DD^T$ for some matrix D , then we have $\bar{x}^T A \bar{x} \geq 0$ for any column vector \bar{x} .

The scalar $\bar{x}^T A \bar{x}$ can be shown to be equal to $\|D^T \bar{x}\|^2$.

6. Show that the matrix product AB remains unchanged if we scale the i th column of A and the i th row of B by respective factors that are inverses of each other.

The idea is to express the matrix multiplication as the sum of outer-products of columns of A and rows of B .

$$AB = \sum_k \bar{A}_k \bar{B}_k$$

Here, \bar{A}_k is the k th column of A and \bar{B}_k is the k th row of B . Note that the expression on the right does not change if we multiply \bar{A}_i by α and divide \bar{B}_i by α . Each component of the sum remains unchanged including the i th component, where the scaling factors cancel each other out.

7. Show that any matrix product AB can be expressed in the form $A' \Delta B'$, where A' is a matrix in which the sum of the squares of the entries in each column is 1, B' is a matrix in which the sum of the squares of the entries in each row is 1, and Δ is an appropriately chosen diagonal matrix with nonnegative entries on the diagonal.

After expressing the matrix product as the sum of outer-products, we can scale each vector in the outer-product to unit-norm, while pulling out a scalar multiple for the outer-product component. The matrices A' and B' contain these normalized vectors, whereas Δ contains these scalar multiples. In other words, consider the case, where we have the product in the following form using the k th column \bar{A}_k of A and the k th row \bar{B}_k of B :

$$AB = \sum_k \bar{A}_k \bar{B}_k$$

One can express this matrix product in the following form:

$$AB = \sum_k \underbrace{\|\bar{A}_k\| \|\bar{B}_k\|}_{\delta_{kk}} \frac{\bar{A}_k}{\|\bar{A}_k\|} \frac{\bar{B}_k}{\|\bar{B}_k\|}$$

We create a diagonal matrix Δ in which the k th diagonal entry is δ_{kk} and then create A' and B' as the normalized versions of A and B , respectively.

8. Discuss how a permutation matrix can be converted to the identity matrix using at most d elementary row operations of a single type. Use this fact to express A as the product of at most d elementary matrix operators.

Only row interchange operations are required to convert it to the identity matrix. In particular, in the i th iteration, we interchange the i th row of A with whatever row contains the i th row of the identity matrix. A permutation matrix will always contain such a row. This matrix can be represented as the product of at most d elementary row interchange operators by treating each interchange operation as a matrix multiplication.

9. Suppose that you reorder all the columns of an invertible matrix A using some random permutation, and you know A^{-1} for the original matrix. Show how you can (simply)

compute the inverse of the reordered matrix from A^{-1} without having to invert the new matrix from scratch. Provide an argument in terms of elementary matrices.

All the rows of A^{-1} are interchanged using exactly the same permutation as the columns of A are permuted. This is because if P is the permutation matrix that creates AP , then $P^T A^{-1}$ is the inverse of AP . However, P^T performs exactly the same reordering on the rows of A as P performs on the columns of A .

10. Suppose that you have approximately factorized an $n \times d$ matrix D as $D \approx UV^T$, where U is an $n \times k$ matrix and V is a $d \times k$ matrix. Show how you can derive an infinite number of alternative factorizations $U'V'^T$ of D , which satisfy $UV^T = U'V'^T$.

Let P be any invertible matrix of size $k \times k$. Then, we set $U' = UP$, and $V' = V(P^{-1})^T$. It can be easily shown that $UV^T = U'V'^T$.

11. Either prove each of the following statements or provide a counterexample:

- (a) The order in which you apply two elementary row operations to a matrix does not affect the final result.
- (b) The order in which you apply an elementary row operation and an elementary column operation does not affect the final result.

It is best to think of these problems in terms of elementary matrix operations.

(a) If you start with the matrix A , then the two successive row operations corresponding to matrices E_1 and E_2 create the matrix $E_2 E_1 A$. Note that matrix multiplication is not commutative and this is not the same as $E_1 E_2 A$. For example, rotation matrices do not commute with scaling matrices. Scaling the first row by 2 followed by interchanging the first and second rows creates a different result than the one obtained by reversing these operations.

(b) In this case, if the row and column operators are E_r and E_c , the final result is $E_r A E_c$. Because of the associativity of matrix multiplication, $(E_r A) E_c$ and $E_r (A E_c)$ are the same. The result follows that the order does not matter.

12. Discuss why some power of a permutation matrix is always the identity matrix.

There are a finite number of permutations of a sequence. Therefore, after some number k of repeated permutations by P , the sequence will be repeated. In other words we have $P^k = I$.

13. Consider the matrix polynomial $\sum_{i=0}^t a_i A^i$. A straightforward evaluation of this polynomial will require $O(t^2)$ matrix multiplications. Discuss how you can reduce the number of multiplications to $O(t)$ by rearranging the polynomial.

The matrix polynomial can be written as $a_0 I + A(\sum_{i=1}^t a_i A^{i-1})$. This can be further expanded as follows:

$$\begin{aligned} a_0 I + A\left(\sum_{i=1}^t a_i A^{i-1}\right) &= a_0 I + A(a_1 I + A\left(\sum_{i=2}^t a_i A^{i-2}\right)) \\ &= a_0 I + A(a_1 I + A(a_2 I + A\left(\sum_{i=3}^t a_i A^{i-2}\right))) \end{aligned}$$

Using this type of expansion recursively, one can obtain the desired result.

14. Let $A = [a_{ij}]$ be a 2×2 matrix with $a_{12} = 1$, and 0s in all other entries. Show that $A^{1/2}$ does not exist even after allowing complex-valued entries.

Suppose that such a matrix exists. If the entries of the 2×2 matrix $A^{1/2}$ listed row-wise are a , b , c , and d , then we obtain the following system of equations:

$$a^2 + bc = 0$$

$$cb + d^2 = 0$$

$$ab + bd = 1$$

$$ca + dc = 0$$

Using the first two equations, we obtain $a^2 - d^2 = 0$, which means either $a = -d$ or $a = d$. Note that $a = -d$ is not possible because the third equation would be violated. Using $a = d$, we can eliminate d to obtain the following system:

$$a^2 + bc = 0$$

$$2ab = 1$$

$$ac = 0$$

Since ab is nonzero and ac is zero, it means that a cannot be zero, and c is zero. However, if c is zero, then the first equation implies that a is zero as well. Therefore, we reach a contradiction.

15. **Parallelogram law:** The parallelogram law states that the sum of the squares of the sides of a parallelogram is equal to the sum of the squares of its diagonals. Write this law as a vector identity in terms of vectors \vec{A} and \vec{B} . Now use vector algebra to show why this vector identity must hold.

The identity is as follows:

$$2\|\vec{A}\|^2 + 2\|\vec{B}\|^2 = \|\vec{A} - \vec{B}\|^2 + \|\vec{A} + \vec{B}\|^2$$

One can expand the right-hand side by using dot products, and then apply the distributive property to show that it is equal to the left-hand side.

$$\|\vec{A} - \vec{B}\|^2 + \|\vec{A} + \vec{B}\|^2 = \vec{A} \cdot \vec{A} - 2\vec{A} \cdot \vec{B} + \vec{B} \cdot \vec{B} + \vec{A} \cdot \vec{A} + 2\vec{A} \cdot \vec{B} + \vec{B} \cdot \vec{B}$$

After canceling out the terms involving $\vec{A} \cdot \vec{B}$ and consolidating others, we get the desired result.

16. Write the first four terms of the Taylor expansion of the following univariate functions about $x = a$: (i) $\log(x)$; (ii) $\sin(x)$; (iii) $1/x$; (iv) $\exp(x)$.

$$(i) \log(a) + (x-a)/a - (x-a)^2/(2a^2) + (x-a)^3/(3a^3)$$

$$(ii) \sin(a) + (x-a)\cos(a) - \frac{(x-a)^2}{2}\sin(a) - \frac{(x-a)^3}{6}\cos(a)$$

$$(iii) \frac{1}{a} - \frac{(x-a)}{a^2} + \frac{(x-a)^2}{a^3} - \frac{(x-a)^3}{a^4}$$

$$(iv) \exp(a) + (x-a)\exp(a) + \frac{(x-a)^2}{2}\exp(a) + \frac{(x-a)^3}{6}\exp(a)$$

17. Use the multivariate Taylor expansion to provide a quadratic approximation of $\sin(x+y)$ in the vicinity of $[x, y] = [0, 0]$. Confirm that this approximation loses its accuracy with increasing distance from the origin.

The Taylor expansion is as follows:

$$[1, 1][x, y]^T + [x, y] \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} [x, y]^T$$

The resulting approximation is $x + y$. Note that if $x = \pi/200$ and $y = \pi/200$ radians, which are small angles, we have $\sin(x + y) = 0.031410 \approx \pi/200 + \pi/200$. However, if we choose large values of x and y like $x = y = \pi$, then $\sin(x + y) = 0 \ll \pi + \pi$.

- 18.** Consider a case where a $d \times k$ matrix P is initialized by setting all values randomly to either -1 or $+1$ with equal probability, and then dividing all entries by \sqrt{d} . Discuss why the columns of P will be (roughly) mutually orthogonal for large values of d of the order of 10^6 . This trick is used frequently in machine learning for rapidly generating the random projection of an $n \times d$ data matrix D as $D' = DP$.

The dot product between any pair of columns has mean of 0 and standard deviation of $1/\sqrt{d}$, since it is the sum of d iid random variables from the Bernoulli distribution. For large values of d like 10^6 , the standard deviation will be of the order of 10^{-3} , and the distribution will be close to normal. Since most of the density of normal distributions is captured between ± 3 standard deviations, it means that the cosine of the angle between each pair of columns will be between -0.003 and 0.003 with high probability. This means that the vectors are very nearly orthogonal. In particular, the pairwise angles will lie between 89.83° and 90.17° with high probability.

- 19.** Consider the perturbed matrix $A_\epsilon = A + \epsilon B$, where the value of ϵ is small and A, B are $d \times d$ matrices. Show the following approximation:

$$A_\epsilon^{-1} \approx A^{-1} - \epsilon A^{-1} B A^{-1}$$

This approximation is useful when A^{-1} is already known.

$$\begin{aligned} A_\epsilon^{-1} &= (A + \epsilon B)^{-1} = [A(I + \epsilon A^{-1} B)]^{-1} = (I + \epsilon A^{-1} B)^{-1} A^{-1} \\ &= (I - \epsilon A^{-1} B + \epsilon^2 (A^{-1} B)^2 - \dots) A^{-1} \approx A^{-1} - \epsilon A^{-1} B A^{-1} \end{aligned}$$

One can verify that the product of A_ϵ with $(A^{-1} - \epsilon A^{-1} B A^{-1})$ differs from the identity matrix by a term dependent on ϵ^2 , which is assumed to be negligible. The approach is particularly efficient when B is very sparse, such as when it contains a small number of nonzero columns.

- 20.** Suppose that you have a 5×5 matrix A , in which the rows/columns correspond to people in a social network in the order John, Mary, Jack, Tim, and Robin. The entry (i, j) corresponds to the number of times person i sent a message to person j . Define a matrix P , so that PAP^T contains the same information, but with the rows/columns in the order Mary, Tim, John, Robin, and Jack.

The permutation matrix is P as follows:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

21. Suppose that the vectors \bar{x} , \bar{y} and $\bar{x} - \bar{y}$ have lengths 2, 3, and 4, respectively. Find the length of $\bar{x} + \bar{y}$ using only vector algebra (and no Euclidean geometry).

This follows from the parallelogram law discussed in an earlier exercise. The corresponding length is $\sqrt{2^2 + 2^2 + 3^2 + 3^2 - 4^2} = \sqrt{10}$.

22. Show that the inverse of a symmetric matrix is symmetric.

Suppose that A is symmetric and B is the inverse of A . Then, we have $AB = BA = I$. Taking the transpose, we obtain $(AB)^T = (BA)^T = I$. Using symmetry of A , this is the same as saying that $B^T A = AB^T = I$. In other words, B^T is the inverse of A as well. However, since the inverse is unique, we must have $B^T = B$. Alternatively, we can use $AB = AB^T = I$ to assert that $A(B - B^T) = 0$. Left multiplying by B , we get $(BA)(B - B^T) = 0$. Since $BA = I$, we have $B - B^T = 0$. In other words, $B = B^T$.

23. Let A_1, A_2, \dots, A_d be $d \times d$ matrices that are strictly upper triangular. Then, the product of A_1, A_2, \dots, A_d is the zero matrix.

Let B_i be given by $A_1 A_2 \dots A_i$. It can be shown inductively that B_i is also strictly triangular, but with at least i zero rows and columns. Therefore, B_d will be the zero matrix.

24. **Apollonius's identity:** Let ABC be a triangle, and AD be the median from A to BC . Show the following using only vector algebra and no Euclidean geometry:

$$AB^2 + AC^2 = 2(AD^2 + BD^2)$$

You will get the simplest algebra by orienting your triangle properly with respect to the origin.

Put the vertex of A at \bar{a} and D as the origin. Then the vertices of B and C are \bar{b} and $\bar{c} = -\bar{b}$. Then, the identity reduces to showing the following:

$$\|\bar{a} - \bar{b}\|^2 + \|\bar{a} + \bar{b}\|^2 = 2(\|\bar{a}\|^2 + \|\bar{b}\|^2)$$

This is easy to show using dot products. In fact, the Apollonius identity reduces to the parallelogram law in this case!

25. **Sine law:** Express the sine of the interior angle between \bar{a} and \bar{b} (i.e., the angle not greater than 180 degrees) purely in terms of $\bar{a} \cdot \bar{a}$, $\bar{b} \cdot \bar{b}$, and $\bar{a} \cdot \bar{b}$. You are allowed to use $\sin^2(x) + \cos^2(x) = 1$. Consider a triangle, two sides of which are the vectors \bar{a} and \bar{b} . The opposite angles to these vectors are A and B , respectively. Show the following using only vector algebra and no Euclidean geometry:

$$\frac{\|\bar{a}\|}{\sin(A)} = \frac{\|\bar{b}\|}{\sin(B)}$$

The sine of the angle is $\sqrt{1 - \left[\frac{\bar{a} \cdot \bar{b}}{\|\bar{a}\| \|\bar{b}\|} \right]^2}$. This is essentially obtained using $\sin(x) = \sqrt{1 - \cos^2(x)}$. Since we are only looking for interior angles, the sine is a positive quantity. Therefore, we can ignore the possibility $\sin(x) = -\sqrt{1 - \cos^2(x)}$.

The angle A is formed between vectors $\bar{a} - \bar{b}$ and $\bar{0} - \bar{b}$. The angle B is formed between vectors $\bar{b} - \bar{a}$ and $\bar{0} - \bar{a}$. Therefore, we obtain the following for the first ratio:

$$\begin{aligned}\frac{\|\bar{a}\|}{\sin(A)} &= \frac{\|\bar{a}\|}{\sqrt{1 - \left[\frac{-\bar{b} \cdot (\bar{a} - \bar{b})}{\|\bar{b}\| \|\bar{a} - \bar{b}\|} \right]^2}} \\ &= \frac{\|\bar{a}\| \|\bar{b}\| \|\bar{a} - \bar{b}\|}{\sqrt{\|\bar{a}\|^2 \|\bar{b}\|^2 - (\bar{a} \cdot \bar{b})^2}}\end{aligned}$$

The second expression is obtained by applying the same approach to the the angle B:

$$\begin{aligned}\frac{\|\bar{b}\|}{\sin(B)} &= \frac{\|\bar{b}\|}{\sqrt{1 - \left[\frac{-\bar{a} \cdot (\bar{b} - \bar{a})}{\|\bar{a}\| \|\bar{b} - \bar{a}\|} \right]^2}} \\ &= \frac{\|\bar{a}\| \|\bar{b}\| \|\bar{b} - \bar{a}\|}{\sqrt{\|\bar{a}\|^2 \|\bar{b}\|^2 - (\bar{a} \cdot \bar{b})^2}}\end{aligned}$$

Note that the two expressions are the same because we have $\|\bar{a} - \bar{b}\| = \|\bar{b} - \bar{a}\|$. This proves the result.

- 26. Trigonometry with vector algebra:** Consider a unit vector $\bar{x} = [1, 0]^T$. The vector \bar{v}_1 is obtained by rotating \bar{x} counter-clockwise at angle θ_1 , and \bar{v}_2 is obtained by rotating \bar{x} clockwise at an angle θ_2 . Use the rotation matrix to obtain the coordinates of unit vectors \bar{v}_1 and \bar{v}_2 . Use this setup to show the following well-known trigonometric identity:

$$\cos(\theta_1 + \theta_2) = \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2)$$

On using the rotation matrix, one obtains the vectors $[\cos(\theta_1), \sin(\theta_1)]$ and $[\cos(\theta_2), -\sin(\theta_2)]$ as follows:

$$\begin{bmatrix} \cos(\theta_1) & -\sin(\theta_1) \\ \sin(\theta_1) & \cos(\theta_1) \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \cos(-\theta_2) & -\sin(-\theta_2) \\ \sin(-\theta_2) & \cos(-\theta_2) \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The value of $\cos(\theta_1 + \theta_2)$ is the cosine of the angle between the vectors, which also be algebraically obtained by using the dot product between the two rotated versions of $[1, 0]^T$. As a result, we algebraically obtain the following dot product:

$$\cos(\theta_1 + \theta_2) = \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2)$$

- 27. Coordinate geometry with matrix algebra:** Consider the two lines $y = 3x + 4$ and $y = 5x + 2$ in the 2-dimensional plane. Find the intersection of the two lines by writing the equations in the following form for appropriately chosen A and \bar{b} :

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \bar{b}$$

Find the intersection coordinates (x, y) of the two lines by inverting matrix A.

The linear equation is as follows:

$$\begin{bmatrix} -3 & 1 \\ -5 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

One can show that the inverse of the matrix is as follows:

$$A^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 5 & -3 \end{bmatrix}$$

On computing $A^{-1}\bar{b}$, one obtains $x = 1$ and $y = 7$.

- 28.** Use the matrix inversion lemma to invert a 10×10 matrix with 1s in each entry other than the diagonal entries which contain the value 2.

The matrix can be written as $I + \bar{u}\bar{u}^T$, where \bar{u} is a column vector of 1s. So we obtain the following:

$$(I + \bar{u}\bar{u}^T)^{-1} = I - \frac{\bar{u}\bar{u}^T}{1 + 10} = I - \frac{\bar{u}\bar{u}^T}{11}$$

In other words, we will have 10/11 on each diagonal entry and $-1/11$ on all other entries in the inverse.

- 29. Solid geometry with vector algebra:** Consider the origin-centered hyperplane in 3-dimensional space that is defined by the equation $z = 2x + 3y$. This equation has infinitely many solutions, all of which lie on the plane. Find two solutions that are not multiples of one another and denote them by the 3-dimensional column vectors \bar{v}_1 and \bar{v}_2 , respectively. Let $V = [\bar{v}_1, \bar{v}_2]$ be a 3×2 matrix with columns \bar{v}_1 and \bar{v}_2 . Geometrically describe the set of all vectors that are linear combinations of \bar{v}_1 and \bar{v}_2 with real coefficients c_1 and c_2 :

$$\mathcal{V} = \left\{ V \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} : c_1, c_2 \in \mathcal{R} \right\}$$

Now consider the point $[x, y, z]^T = [2, 3, 1]^T$, which does not lie on the above hyperplane. We want to find a point \bar{b} on the hyperplane for which \bar{b} is as close to $[2, 3, 1]^T$ as possible. How is the vector $\bar{b} - [2, 3, 1]^T$ geometrically related to the hyperplane? Use this fact to show the following condition on \bar{b} :

$$V^T \left(\bar{b} - \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Find a way to eliminate the 3-variable vector \bar{b} from the above equation and replace with the 2-variable vector $\bar{c} = [c_1, c_2]^T$ instead. Substitute numerical values for entries in V and find \bar{c} and \bar{b} with a 2×2 matrix inversion.

Two such solutions are $\bar{v}_1 = [1, 0, 2]^T$ and $\bar{v}_2 = [0, 1, 3]^T$. The set of all points in \mathcal{V} defines the set of all vectors on the surface of the hyperplane corresponding to the equation. The above condition holds because the point \bar{b} must be normal to the hyperplane in order to be the closest point to it. Therefore, all vectors on the surface

of this origin-centered hyperplane must be normal to the line joining \bar{b} and the line $[2, 3, 1]^T$. In other words, we have:

$$\bar{v}_1 \cdot \left(\bar{b} - \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \right) = 0$$

$$\bar{v}_2 \cdot \left(\bar{b} - \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \right) = 0$$

Stacking up these conditions in matrix form, we get the condition in the statement of the problem. Furthermore, one can instead set $\bar{b} = V\bar{c}$ and solve for \bar{c} instead of \bar{b} . This leads to the following condition:

$$V^T V \bar{c} = V^T \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

Simplifying, we obtain:

$$\begin{bmatrix} 5 & 6 \\ 6 & 10 \end{bmatrix} \bar{c} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

Inverting \bar{c} , we obtain the following:

$$\bar{c} = \frac{1}{14} \begin{bmatrix} 10 & -6 \\ -6 & 5 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

In other words, we have $\bar{c} = \frac{1}{7}[2, 3]^T$. One can now derive \bar{b} as $V\bar{c}$, which is as follows;

$$\bar{b} = \frac{1}{7} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 2 \\ 3 \\ 13 \end{bmatrix}$$

- 30.** Let A and B be two $n \times d$ matrices. One can partition them columnwise as $A = [A_1, A_2]$ and $B = [B_1, B_2]$, where A_1 and B_1 are $n \times k$ matrices containing the first k columns of A and B , respectively, in the same order. Let A_2 and B_2 contain the remaining columns. Show that the matrix product AB^T can be expressed as follows:

$$AB^T = A_1 B_1^T + A_2 B_2^T$$

As discussed in the chapter, each matrix multiplication AB^T can be expressed as the sum of the outer-products between columns of A and B . One can do the same for both A_1, B_1 and A_2, B_2 . In each case, the products turn out to be the same.

- 31. Matrix centering:** In machine learning, a common centering operation of an $n \times n$ similarity matrix S is the update $S \leftarrow (I - U/n)S(I - U/n)$, where U is an $n \times n$ matrix of 1s. How would you use the associative property of matrix multiplication to implement this operation efficiently.

The matrix U can be expressed as $\bar{1}\bar{1}^T$, where $\bar{1}$ is a column vector of 1s. We can write the above matrix multiplication as follows:

$$(I - U/n)S(I - U/n) = S - US/n - SU/n + USU/n^2$$

The above expression requires the computation of US , SU , and USU , aside from some cheap matrix addition operations. The matrix US can be computed as $\bar{1}(\bar{1}^T S)$, where the bracketing tells us about the ordering of the multiplications. Furthermore, the matrix SU can be computed as $(S\bar{1})\bar{1}^T$, where the bracketing tells us about the ordering of the multiplications. The final matrix is USU , which can be written as $\bar{1} \underbrace{[\bar{1}^T (S\bar{1})]}_{\text{Scalar}} \bar{1}^T$. All of these operations are simple matrix-to-vector multiplications, which can be done very cheaply.

- 32. Energy preservation in orthogonal transformations:** Show that if A is an $n \times d$ matrix and P is a $d \times d$ orthogonal matrix, then we have $\|AP\|_F = \|A\|_F$.

One can use the relationship with the trace.

$$\|AP\|_F^2 = \text{tr}(AP(AP)^T) = \text{tr}(A(PP^T)A^T) = \text{tr}(AA^T) = \|A\|_F^2$$

- 33. Tight sub-multiplicative case:** Suppose that \bar{u} and \bar{v} are column vectors (of not necessarily the same dimensionality). Show that the matrix $\bar{u}\bar{v}^T$ created from the outer product of \bar{u} and \bar{v} has Frobenius norm of $\|\bar{u}\| \|\bar{v}\|$.

One can again use the properties of the trace. The squared Frobenius norm is as follows:

$$\text{tr}(\bar{u}[\bar{v}^T \bar{v}]\bar{u}^T) = \|\bar{v}\|^2 \text{tr}(\bar{u}\bar{u}^T) = \|\bar{v}\|^2 \text{tr}(\bar{u}^T \bar{u}) = \|\bar{v}\|^2 \|\bar{u}\|^2$$

It is important to note that we were able to pull out $\|\bar{v}\|^2$ from the trace in one of the above steps, because the expression $\bar{v}^T \bar{v}$ is a scalar, and simply scales up all the matrix entries uniformly.

- 34. Frobenius orthogonality and Pythagorean theorem:** Two $n \times d$ matrices A and B are said to be Frobenius orthogonal if the sum of entry-wise products of their corresponding elements are zero [i.e., $\text{tr}(AB^T) = 0$]. Show the following:

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2$$

The squared Frobenius norm can be expressed in terms of the trace as follows:

$$\begin{aligned} \|A + B\|_F^2 &= \text{tr}((A + B)(A + B)^T) = \text{tr}(AA^T) + \text{tr}(BB^T) + \underbrace{\text{tr}(AB^T) + \text{tr}(BA^T)}_0 \\ &= \|A\|_F^2 + \|B\|_F^2 \end{aligned}$$

35. Let \bar{x} and \bar{y} be two orthogonal column vectors of dimensionality n . Let \bar{a} and \bar{b} be two arbitrary d -dimensional column vectors. Show that the outer products $\bar{x}\bar{a}^T$ and $\bar{y}\bar{b}^T$ are Frobenius orthogonal (see Exercise 34 for definition of Frobenius orthogonality).

The sum of the entry-wise products in the two matrices is given by $\text{tr}((\bar{x}\bar{a}^T)^T(\bar{y}\bar{b}^T))$. On simplifying, one will obtain $\bar{x}^T\bar{y}$ in the middle, which evaluates to 0. As a result all entries in the product matrix will be zero. Therefore, the trace will be zero as well.

36. Suppose that a sequence of row and column operations is performed on a matrix. Show that as long as the ordering among row operations and the ordering among column operations is maintained, changing the ordering between row and column operations does not make any difference to the final result.

If the row operations have operator matrices in the order $R_1 \dots R_k$ and the column operations have operator matrices in the order $C_1 \dots C_k$, the overall transformation of A can be expressed as $R_k \dots R_1 A C_1 \dots C_k$. Note that we can group these operator matrices in any way we like because of the associative property of matrix multiplication.

37. Show that any orthogonal upper-triangular matrix is a diagonal matrix.

This result can be shown by case-by-case analysis of matrix U . First note that the diagonal elements of the product UU^T are the products of the diagonal elements of U and U^T . Therefore, all diagonal elements of U are nonzero, or else the product of the two can never have 1s on the diagonal elements in order to create the identity matrix. In fact, the values on the diagonal will be 1 and -1. Second, note that dot products between pairs of columns of U are zero. Since the dot product between the first and second column of U , which is $u_{11}u_{12}$, is zero, it means that the u_{12} must be zero. This is because u_{11} is guaranteed to be non-zero. Next, we take the dot product of the third column with the second column and first column (in that order) to show that the off-diagonal elements of the third column are zeros. In general, for the r th column, we take the dot product with the $(r-1)$ th, $(r-2)$ th, and so on in that order in order to show that successive off-diagonal elements are zeros.

Simpler proof: The inverse of an upper-triangular matrix is also upper triangular. However, the inverse of an orthogonal matrix is its transpose which happens to be lower triangular. Therefore, we obtain a matrix that is both upper triangular and lower triangular. This can happen only when the matrix is diagonal.

38. Consider a set of vectors $\bar{x}_1 \dots \bar{x}_n$, which are known to be unit normalized. You do not have access to the vectors but you are given all pairwise squared Euclidean distances in the $n \times n$ matrix Δ . Discuss why you can derive the $n \times n$ pairwise dot product matrix by adding 1 to each entry of the matrix $-\frac{1}{2}\Delta$.

When the vectors are unit normalized, the dot product and Euclidean distance can be derived from one another without any additional information about placement of points. Consider the two unit vectors \bar{x} and \bar{y} . Then, we have the following:

$$\text{Euclidean}(\bar{x}, \bar{y})^2 = (\bar{x} - \bar{y}) \cdot (\bar{x} - \bar{y}) = \underbrace{\bar{x} \cdot \bar{x} + \bar{y} \cdot \bar{y}}_{1+1} - 2\bar{x} \cdot \bar{y} = 2 - 2\bar{x} \cdot \bar{y}$$

This implies that the dot product is obtained by multiplying the squared Euclidean distance with -0.5 and then adding 1. The result follows.

39. We know that every matrix commutes with its inverse. We want to show a generalization of this result. Consider the polynomial functions $f(A)$ and $g(A)$ of matrix A , so that $f(A)$ is invertible. Show the following commutative property:

$$[f(A)]^{-1}g(A) = g(A)[f(A)]^{-1}$$

We know that $f(A)g(A) = g(A)f(A)$. This is because polynomials of the same matrix commute. In fact, a large number of useful identities in matrix algebra, such as the push-through identity, indirectly depend on this fact. Then, by both right multiplying and left multiplying with $f(A)^{-1}$, we get the desired result.

40. Given an example of a 2×2 matrix A and a polynomial function $f(\cdot)$, so that A is invertible but $f(A)$ is not invertible. Given an example of a matrix A so that A is not invertible but $f(A)$ is invertible. Note that the constant term in the polynomial corresponds to a multiple of the identity matrix.

Let A be the identity matrix, and $f(A) = A - I$, which is the zero matrix. Then, A is invertible but $f(A)$ is not. Also let A be the 2×2 matrix with a single 1 for the entry a_{11} and zeros in all other entries. Let $f(A) = A + I$. Then A is not invertible, but $f(A)$ is invertible.

41. Let A be a rectangular matrix and $f(\cdot)$ be a polynomial function. Show that $A^T f(AA^T) = f(A^T A)A^T$. Now suppose that both $f(AA^T)$ and $f(A^T A)$ are invertible. Show the following result:

$$[f(A^T A)]^{-1}A^T = A^T[f(AA^T)]^{-1}$$

Interpret the push-through identity as a special case of this result.

One can write $f(AA^T)$ as follows:

$$f(AA^T) = \sum_{i=0}^n c_i (AA^T)^i$$

Left-multiplying both sides with A^T , we obtain the following:

$$A^T f(AA^T) = A^T \sum_{i=0}^n c_i (AA^T)^i = \left[\sum_{i=0}^n c_i (A^T A)^i \right] A^T = f(A^T A)A^T$$

Therefore, we obtain the first result. Now consider the case where we left multiply this identity with $f(A^T A)^{-1}$ and right-multiply this identity with $f(AA^T)^{-1}$, we obtain the second result. The push-through identity is a special case of this result by setting $f(AA^T) = AA^T + \lambda I$.

42. Discuss why one cannot generalize the formula for the scalar binomial expansion $(a + b)^n$ to the matrix expansion $(A + B)^n$. Also discuss why generalization is possible in cases where $B = f(A)$ for some polynomial function $f(\cdot)$.

This formula cannot be generalized because of non-commutativity of matrices. For example, $(A + B)^2$ is $A^2 + AB + BA + B^2$, and the expressions for AB and BA cannot be consolidated. The problem increases with increasing value of n . In the case, where

$B = f(A)$, the matrices A and B commute. Therefore, the terms can be consolidated and the formula continues to hold. In fact, the binomial expansion holds for matrices if and only if the two matrices in the expansion commute. This does happen quite frequently in many practical settings.

43. Suppose that A is a $d \times d$ matrix satisfying $A^4 = 0$. Derive an algebraic expression for $(I + A)^{-1}$ as a matrix polynomial in A .

The formula is $I - A + A^2 - A^3$. This result can be obtained by using the subsection on the result on the computation of $(I + A)^{-1}$, which is an infinite series. Note that multiplying with $I + A$ yields $I - A^4$, which is the identity matrix, since we have $A^4 = 0$.

44. Compute the inverse of the following triangular matrix by expressing it as the sum of two matrices:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 3 & 1 \end{bmatrix}$$

One can express the matrix as $(I + A)$, since the values along the diagonal are 1s. Here, A is a strictly triangular matrix satisfying the nilpotency condition $A^3 = 0$. Therefore, the inverse is given by $(I + A)^{-1} = I - A + A^2$. This value can be shown to be the following matrix:

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 5 & -3 & 1 \end{bmatrix}$$

45. Express a $d \times d$ matrix M of 1s as the outer product of two d -dimensional vectors. Use the matrix inversion lemma to compute an algebraic expression for $(I + M)^{-1}$.

Using the matrix inversion lemma, one obtains the expression $I - M/(n + 1)$.

46. Show that if A and B commute, the matrix polynomials $f(A)$ and $g(B)$ commute as well.

One can show the result by using the distributive property of matrix multiplication, then swapping the order of A and B in the individual terms, and then regrouping them back.

47. Show that if invertible matrices A and B commute, A^k and B^s commute for any integers $k, s \in [-\infty, \infty]$. Show that the previous exercise holds for “polynomials” with both positive and negative integer exponents included.

This exercise can be shown by using case-wise analysis. First, note that if either k or s is zero, the proof becomes trivial. Therefore, we will only consider nonzero k and s . The following are the cases:

Case I $k > 0, s > 0$: In this case, we have $A^k B^s = B A^k B^{s-1}$. This is because we have:

$$A^k B^s = A^{k-1} (AB) B^{s-1} = A^{k-1} (BA) B^{s-1} = A^{k-2} (BA) AB^{s-1} = \dots = B A^k B^{s-1}$$

The overall effect is to pull a single matrix B from the end of the expression to the front of the expression. We repeatedly pull each and every B to the front

using the commutative property. By repeating the process for each of matrices B , we get $A^k B^s = B^s A^k$.

Case II $k > 0, s < 0$. Note that $A^k B^{-s} = B^{-s} A^k$ using the argument of Case I. Post-multiplying and pre-multiplying the above equation with B^s , we get the desired result.

Case III $k < 0, s > 0$. This case is exactly analogous to Case II.

Case IV $k < 0, s < 0$. In this case, we have $A^{-k} B^{-s} = B^{-s} A^{-k}$ by using the argument of Case I. Pre-multiplying and post-multiplying both sides with B^s , we get $B^s A^{-k} = A^{-k} B^s$. Then, pre-multiplying and post-multiplying both sides with A^k we get $A^k B^s = B^s A^k$, which is the desired result.

The proof for the extended definition of polynomials (with negative exponents included) is similar to that of the previous exercise. We can use the distributive property of matrix multiplication and then swap the order of A and B in the multiplicative terms.

48. Let $U = [u_{ij}]$ be an upper-triangular $d \times d$ matrix. What are the diagonal entries of $f(U)$ as scalar functions of the matrix entries u_{ij} ?

The diagonal entries of $f(U)$ are $f(\lambda_1) \dots f(\lambda_d)$, where $\lambda_i = u_{ii}$. This because the diagonal entries of U^k can be shown to be $\lambda_1^k \dots \lambda_d^k$. The diagonal entries of the sum of two matrices are the sum of their diagonal entries.

49. **Inverses behave like matrix polynomials:** The Cayley-Hamilton theorem states that a finite-degree polynomial $f(\cdot)$ always exists for any matrix A satisfying $f(A) = 0$. Use this fact to prove that the inverse of A is also a finite-degree polynomial.

The Cayley-Hamilton theorem is introduced in Chapter 3. Although invertible matrices can be shown to have some special properties of the polynomial, such as a nonzero constant term, we will not make this assumption in this proof, since these results have not been introduced yet. We will only assume that an inverse of A exists. Let the smallest degree in the polynomial $f(A)$ be r (which is guaranteed to be 0 in the case of invertible matrices but we do not make this assumption). We multiply $f(A)$ with A^{-r-1} to obtain the following:

$$A^{-r-1} f(A) = 0$$

The above matrix polynomial can be expressed in the form $cA^{-1} + g(A)$ for some matrix polynomial $g(A)$ and non-zero coefficient c . Therefore, the inverse of A is simply expressed as $-g(A)/c$.

50. Derive the inverse of a 3×3 row addition operator by inverting the sum of matrices.

A 3×3 row addition operator matrix can be expressed as $(I + C)$ where C is a matrix containing a single off-diagonal entry. We use the following result:

$$(I + C)^{-1} = I - C + C^2 - C^3 + \dots$$

Since C contains only a single off-diagonal entry, it is a nilpotent matrix satisfying $C^2 = 0$. Therefore, the inverse is given by the following:

$$(I + C)^{-1} = I - C$$

51. For any non-invertible matrix A , show that the infinite summation $\sum_{k=0}^{\infty} (I - A)^k$ cannot possibly converge to a finite matrix. Give two examples to show that if A is invertible, the summation might or might not converge.

Suppose that the summation does converge to a finite matrix. Then, we know that $(I - A)^k$ tends to the zero matrix as k becomes large. Now multiplying the summation with A , we obtain the following:

$$A \sum_{k=0}^{\infty} (I - A)^k = [I - (I - A)] \sum_{k=0}^{\infty} (I - A)^k$$

One can use the distributive property of matrix multiplication to obtain the following:

$$A \sum_{k=0}^{\infty} (I - A)^k = \sum_{k=0}^{\infty} [(I - A)^k - (I - A)^{k+1}]$$

This summation has adjacent terms that cancel each other out, and the trailing terms converge to zero as k gets very large. Therefore, one is left with only the first term, which is $(I - A)^0 = I$. In other words, the summation is the inverse of the matrix A . This is a contradiction to the original statement of the problem.

Now, we consider the cases where A is invertible. Choosing $A = I/2$, results in a converging summation. Choosing $A = 3I$ results in a summation that does not converge.

52. The chapter shows that the product, $A_1 A_2 \dots A_k$, of invertible matrices is invertible. Show the converse that if the product $A_1 A_2 \dots A_k$ of square matrices is invertible, each matrix A_i is invertible. Use only the material discussed in this chapter for the proof.

If $A_1 A_2 \dots A_k$ is invertible, there must exist a matrix C , such that the following is true:

$$C A_1 A_2 \dots A_k = I$$

Therefore, the product of $C A_1 A_2 \dots A_{k-1}$ and A_k is I . This is possible only if A_k is invertible. Furthermore, we can show that $A_1 A_2 \dots A_{k-1}$ is invertible, because we have:

$$[A_1 A_2 \dots A_{k-1}] [A_k C] = I$$

What we have achieved is to use the invertibility of $A_1 A_2 \dots A_k$ to show the invertibility of the smaller pieces $A_1 A_2 \dots A_{k-1}$ and A_k . We can then apply the same approach to $A_1 A_2 \dots A_{k-1}$ to show that $A_1 A_2 \dots A_{k-2}$ and A_{k-1} are invertible. This process is repeated to show that each matrix A_i is invertible.

53. Show that if a $d \times d$ diagonal matrix Δ with distinct diagonal entries $\lambda_1 \dots \lambda_d$ commutes with A , then A is diagonal.

Suppose that a non-diagonal entry a_{ij} of A is non-zero. Then, the (i, j) th entry of $A\Delta$ is $a_{ij}\lambda_j$, whereas the (i, j) th entry of ΔA is $\lambda_i a_{ij}$. Since $\lambda_i \neq \lambda_j$, and $a_{ij} \neq 0$, these values cannot be equal. Therefore, we have a contradiction, and it follows that a_{ij} cannot be nonzero for non-diagonal entries. Therefore, A is a diagonal matrix.

54. What fraction of 2×2 binary matrices with 0-1 entries are invertible?

All matrices with zero 1s, one 1s, and four 1s are not invertible. There are $1 + 4 + 1 = 6$ such matrices. All matrices with three 1s are invertible, and there are four such matrices. Among matrices with two 1s, the four matrices with 1s in the same row or column are not invertible, whereas the two others are invertible. Therefore six out of sixteen binary matrices are invertible. The required fraction is $6/16 = 3/8$.

Chapter 2

Linear Transformations and Linear Systems

1. *If we have a square matrix A that satisfies $A^2 = I$, then, is it always the case that $A = \pm I$? Either prove the statement or provide a counterexample.*

This statement is not always true. For example, flipping the sign of only a subset of the diagonal elements of the identity matrix results in a matrix A , which satisfies $A^2 = I$.

2. *Show that the matrices A , AA^T , and $A^T A$ must always have the same rank for any $n \times d$ matrix A .*

First, we will show that the ranks of A and $A^T A$ are the same. If the vector \bar{x} belongs to the right null space of A then we have $A\bar{x} = 0$. Multiplying with A^T we get $A^T A\bar{x} = 0$. In other words, \bar{x} belongs to the right null space of $A^T A$. Similarly, if \bar{x} belongs to the right null space of $A^T A$, one can show that $\|A\bar{x}\|^2 = \bar{x}^T A^T A\bar{x} = 0$. In other words, we have $A\bar{x} = 0$ and therefore \bar{x} is in the null space of A . Therefore, the null spaces of A and $A^T A$ are the same. Since both matrices have d columns (i.e., vectors of length d in rows), the sum of the ranks of the right null space and row space must be d in both cases. Therefore, the row rank is the same in both cases. But the row rank is the same as the matrix rank. In other words, the ranks of A and $A^T A$ must be the same. One can use a similar result to show that the ranks of A^T and AA^T are the same by simply applying the entire argument to A^T instead of A . Since the ranks of A and A^T are the same, the result follows.

3. *Provide a geometric interpretation of A^9 , where A is a 2×2 rotation matrix at a counter-clockwise angle of 60° .*

If a rotation matrix rotates by θ , its k th power rotated by $k\theta$. Therefore, we have a counterclockwise rotation of 540° , which is the same as a 180° rotation.

4. *Consider 6×10 matrices A and B of rank 6. What is the minimum and maximum possible rank of the 6×6 matrix AB^T . Provide examples of both cases.*

The maximum possible rank is AB^T is 6. An example of this case is one in which A has orthonormal rows, and we set $B = A$. The product AB^T will be the 6×6 identity matrix.

The minimum possible rank based on Sylvester's inequality is $6 + 6 - 10 = 2$. In this case, we set A to have 6 orthonormal rows. B is constructed by choosing 2 of its rows from B , and the other four rows to be orthonormal to all the rows in A . The resulting product AB^T will be a diagonal matrix with its first two diagonal entries set to 1, and the remaining entries set to 0.

5. Use each of row reduction and Gram-Schmidt to find basis sets for the span of $\{[1, 2, 1]^T, [2, 1, 1]^T, [3, 3, 2]^T\}$. What are the best-fit coordinates of $[1, 1, 1]^T$ in each of these basis sets? Verify that the best-fit vector is the same in the two cases.

The answer to this question is not unique, and it will depend heavily on the order in which one processes the rows. In the case of row reduction, one possible basis set is $[1, 2, 1]^T$, and $[0, 3, 1]$. In the case of Gram-Schmidt orthogonalization, the basis set is $[1, 2, 1]^T/\sqrt{6}$, $[7, -4, 1]^T/\sqrt{66}$. One can create a 3×2 matrix A using the two columns $[1, 2, 1]^T$, and $[0, 3, 1]$ and compute the coordinates of $[1, 1, 1]^T$ as follows:

$$\bar{x} = (A^T A)^{-1} A^T \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

The resulting coordinates that are the components of vector \bar{x} in $A\bar{x} = \bar{b}$ are $12/11$ and $-4/11$. The best fit $A\bar{x}$ of $[1, 1, 1]^T$ with the use of these coordinates is $[12/11, 12/11, 8/11]^T$.

In the case of Gram-Schmidt orthogonalization, the coordinates are much easier to compute with the use of dot products. Using dot product of the basis vectors with $[1, 1, 1]^T$, we obtain the coordinates as $4/\sqrt{6}$ and $4/\sqrt{66}$. The best fit vector is the transposition of $4*[1, 2, 1]/6 + 4*[7, -4, 1]/66$. This vector evaluates to the transposition of $4 * [18, 18, 12]/66 = [12, 12, 8]/11$.

6. Propose a test using Gram-Schmidt orthogonalization to identify whether two sets of (possibly linearly dependent) vectors span the same vector space.

Let S_1 and S_2 be two sets of vectors. Find a Gram-Schmidt basis for S_1 . Now continuing to apply Gram-Schmidt for each vector in S_2 will always evaluate to the zero vector. Therefore, S_2 is a subspace of S_1 . Now perform this same approach in reversed order by creating a basis for S_2 first. If all vectors in S_1 evaluate to the zero vector, it will show that S_1 is a subspace of S_2 . Combining the two, we get the fact that the two vector spaces are the same.

Another approach: For the first set S_1 , first find an orthonormal basis using Gram-Schmidt. For the second set S_2 , find an orthonormal basis using Gram-Schmidt. First, the dimensionality of both sets must be the same. Second, test if the norm of each vector in S_2 is preserved when represented using coordinates in the orthonormal basis for S_1 . (Basis transformation is easy for orthonormal sets). If this is the case then both sets have the same span. Note that one does not need to test whether the norm of each vector in S_1 is preserved when represented using the orthonormal basis for S_2 . This is already guaranteed based on the two tests that have been done.

7. A $d \times d$ skew symmetric matrix satisfies $A^T = -A$. Show that all diagonal elements of such a matrix are 0. Show that $\bar{x} \in \mathcal{R}^d$ is orthogonal to $A\bar{x}$ if and only if A is skew symmetric. Discuss the difference between this transform and a pure rotation by 90° .

Since $A^T = -A$, we have $a_{ii} = -a_{ii}$ which is possible only when $a_{ii} = 0$. Since the transpose of a scalar is always the same scalar, we have $\bar{x}^T A \bar{x} = \bar{x}^T A^T \bar{x}$. However, this means that $\bar{x}^T A \bar{x} = -\bar{x}^T A \bar{x}$. This means that $\bar{x}^T A \bar{x} = 0$. Therefore, \bar{x} is orthogonal to $A\bar{x}$.

To prove the converse, we assume that $\bar{x}^T A \bar{x} = 0$. We first show that the diagonal entries are zero. We pick a vector \bar{x} that is the same as \bar{e}_i . In such a case, we have $\bar{e}_i^T A \bar{e}_i = a_{ii} = 0$. Next, we choose a vector \bar{x} with 1s in the i th and j th entry and zeros in all other entries. In such a case, we have $0 = \bar{x}^T A \bar{x} = a_{ii} + a_{jj} + a_{ij} + a_{ji} = a_{ij} + a_{ji}$. Therefore, $a_{ij} = -a_{ji}$ and the matrix is skew symmetric.

Note that a rotation matrix preserves the norm of a vector. A skew symmetric matrix might not preserve the norm of a vector.

8. Consider the 4×4 Givens matrix $G_c(2, 4, 90)$. This matrix performs a 90° counter-clockwise rotation of a 4-dimensional column vector in the 2-dimensional projection corresponding to the second and fourth dimensions. Show how to obtain this matrix as the product of two Householder reflection matrices. It is strongly encouraged to think geometrically in order to solve this problem. Is the answer to this question unique?

This problem is first solved for 2×2 matrices, where it can be shown that a 90° counter-clockwise rotation is two reflections. A 90° counter-clockwise rotation is first a reflection across $x = y$ and then a reflection across the Y -axis. One can generalize this idea to 4×4 matrices by performing exactly this operation in that projection. The answer to this question is not unique, as one can first reflect across the X -axis and then reflect across $x = y$ to achieve the same result. As long as the two reflection lines are at an angle of 45° to one another, the approach will work.

9. Repeat Exercise 8 for a Givens matrix that rotates a column vector counter-clockwise for 10° instead of 90° .

The best way of solving this problem is to examine the case of 2×2 matrices. Successive reflection in two lines at an angle of $\theta/2$ causes rotation at an angle of θ . The ordering of the two reflections decides whether the rotation is clockwise or counter-clockwise. One can generalize this result to $d \times d$ Givens matrices by performing the operation in the corresponding 2-dimensional projection.

10. Consider the 5×5 matrices A , B , and C , with ranks 5, 2, and 4, respectively. What is the minimum and maximum possible rank of $(A + B)C$.

The matrix $A + B$ can have minimum rank of 3 and maximum rank of 5. Therefore, maximum rank of $(A + B)C$ is 4 since C has rank 4. The minimum is $3 + 4 - 5 = 2$ using the Sylvester inequality.

11. Solve the following system of equations using the Gaussian elimination procedure discussed in the book:

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$$

Take care to use the same conventions for subtracting rows and columns as discussed in the book. Now use these row operations to create an LU decomposition. Is it possible to perform an LU decomposition of this matrix without the use of a permutation matrix?

On subtracting row 2 from row 3, and row 1 from row 2, we obtain the following:

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3-2 \\ 4-3 \end{bmatrix}$$

Swapping the rows will lead to lower triangular form. The solution by back substitution results in $x_i = 1$ for each i . A permutation matrix is essential to get the LU decomposition because swapping rows is essential.

- 12.** Solve the system of equations in the previous exercise using QR decomposition. Use the Gram-Schmidt method for orthogonalization. Use the QR decomposition to compute the inverse of the matrix if it exists.

The solution to the system of equations is $x_i = 1$ for all i .

- 13.** Why must the column space of matrix AB must be a subspace of the column space of A ? Show that all four fundamental subspaces of A^{k+1} must be the same as that of A^k for some integer k .

This is the case because each column of AB is a linear combination of the columns of A . Note that if A^{k+1} does not have the same column space as A^k , that of A^{k+1} must be a proper subspace of A^k . This can happen for at most d times for a $d \times d$ matrix. Therefore, for some integer k at most equal to d , the condition must be satisfied. Note that if the column space of the matrix does not change, the rank of the row space will not change either. Therefore, the row space will also remain unchanged and all four fundamental subspaces will remain the same.

- 14.** Consider a vector space $\mathcal{V} \subset \mathcal{R}^3$ and two of its possible basis sets $\mathcal{B}_1 = \{[1, 0, 1]^T, [1, 1, 0]^T\}$ and $\mathcal{B}_2 = \{[0, 1, -1]^T, [2, 1, 1]^T\}$. Show that \mathcal{B}_1 and \mathcal{B}_2 are basis sets for the same vector space. What is the dimensionality of this vector space? Now consider a vector $\bar{v} \in \mathcal{V}$ with coordinates $[1, 2]^T$ in basis \mathcal{B}_1 , where the order of coordinates matches the order of listed basis vectors. What is the standard basis representation of \bar{v} ? What are the coordinates of \bar{v} in \mathcal{B}_2 ?

First, note that the vectors in each of \mathcal{B}_1 and \mathcal{B}_2 are linearly independent. Each of the vectors in \mathcal{B}_2 can be expressed in terms of the vectors in \mathcal{B}_1 as follows:

$$\begin{aligned} [0, 1, -1]^T &= [1, 1, 0]^T - [1, 0, 1]^T \\ [2, 1, 1]^T &= [1, 1, 0]^T + [1, 0, 1]^T \end{aligned}$$

Similarly, one can express the vectors in \mathcal{B}_1 in terms of those in \mathcal{B}_2 . Since both sets contain linearly independent vectors, each is the basis of some vector space. Furthermore, since all vectors can be expressed terms of each \mathcal{B}_1 and \mathcal{B}_2 , it implies that these vector spaces are the same. The dimensionality of each vector space is the size of the basis, which is 2.

The vector \bar{b} in the standard basis is as follows:

$$\bar{v} = [1, 0, 1]^T + 2 * [1, 1, 0]^T = [3, 2, 1]^T$$

The coordinates in the second basis are $[1, 3]^T$, and these can be computed using the left-inverse.

15. Find the projection matrix of the following matrix using the QR method:

$$A = \begin{bmatrix} 3 & 6 \\ 0 & 1 \\ 4 & 8 \end{bmatrix}$$

How can you use the projection matrix to determine whether the vector $\bar{b} = [1, 1, 0]^T$ belongs to the column space of A ? Find a solution (or best-fit solution) to $A\bar{x} = \bar{b}$.

The QR decomposition is as follows:

$$A = \begin{bmatrix} 3/5 & 0 \\ 0 & 1 \\ 4/5 & 0 \end{bmatrix} \begin{bmatrix} 5 & 10 \\ 0 & 1 \end{bmatrix}$$

The first matrix in the decomposition is Q . The projection matrix is QQ^T .

$$\begin{bmatrix} 3/5 & 0 \\ 0 & 1 \\ 4/5 & 0 \end{bmatrix} \begin{bmatrix} 3/5 & 0 & 4/5 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 9/25 & 0 & 12/25 \\ 0 & 1 & 0 \\ 12/25 & 0 & 16/25 \end{bmatrix}$$

The projection of the vector should be itself when it does belong to the column space of the matrix. The projection of vector \bar{x} is $P\bar{x}$, where $P = QQ^T$ is the projection matrix. Therefore, the projection of $[1, 1, 0]^T$ is as follows:

$$\begin{bmatrix} 9/25 & 0 & 12/25 \\ 0 & 1 & 0 \\ 12/25 & 0 & 16/25 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 9/25 \\ 1 \\ 12/25 \end{bmatrix}$$

Since the projection of the vector is not itself, it follows that the best fit is the vector on the right-hand side of the above equation.

The solution to the system is $\bar{x} = R^{-1}Q^T\bar{b}$ which is as follows:

$$\begin{bmatrix} 1/5 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3/5 \\ 1 \end{bmatrix} = \begin{bmatrix} -47/25 \\ 1 \end{bmatrix}$$

16. For the problem in Exercise 15, does a solution exist to $A^T\bar{x} = \bar{c}$, where $\bar{c} = [2, 2]^T$? If no solution exists, find the best-fit. If one or more solutions exist, find the one for which $\|\bar{x}\|$ is as small as possible.

Yes, a solution does exist because the column space of A spans all of \mathcal{R}^2 . One example of a solution is $[2/3, -2, 0]^T$. There are an infinite number of solutions, since the columns of A^T are linearly dependent. The right inverse is as follows (which is the same as the Moore-Penrose pseudoinverse):

$$A^+ = \begin{bmatrix} 0.12 & 0 \\ -2 & 1 \\ 0.16 & 0 \end{bmatrix}$$

The most concise solution is $\bar{x} = A^+c$ using the right inverse, and the corresponding coordinate vector is $[0.24, -2, 0.32]^T$.

- 17. Gram-Schmidt with Projection Matrix:** Given a set of $m < n$ linearly independent vectors $\bar{a}_1 \dots \bar{a}_m$ in \mathcal{R}^n , let A_r be the $n \times r$ matrix defined as $A_r = [\bar{a}_1, \bar{a}_2, \dots, \bar{a}_r]$ for each $r \in \{1 \dots m\}$. Show the result that after initializing $\bar{q}_1 = \bar{a}_1$, the unnormalized Gram-Schmidt vectors $\bar{q}_2 \dots \bar{q}_m$ of $\bar{a}_2 \dots \bar{a}_m$ can be computed non-recursively as follows:

$$\bar{q}_{s+1} = [I - A_s(A_s^T A_s)^{-1} A_s^T] \bar{a}_{s+1} = \bar{a}_{s+1} - [P_s \bar{a}_{s+1}] \quad \forall s \in \{1, \dots, m-1\}$$

Here, P_s is the projection matrix derived from A_s .

This result can be shown by proving inductively that \bar{q}_{s+1} is orthogonal to \bar{q}_r for $r \leq s$. We first assume that orthogonality holds between each pair \bar{q}_i and \bar{q}_j for $i, j \leq s$. Now, we will show that \bar{q}_{s+1} is also orthogonal to each \bar{q}_r for all $r \leq s$. First note that P_s can also be rewritten as $Q_s Q_s^T$, where the j th column of Q_s is $\bar{q}_j / \|\bar{q}_j\|$. This is because we can write $A_s = Q_s R$ using the QR decomposition based on the inductive assumption. Substituting for $A_s = Q_s R$ in $P_s = A_s(A_s^T A_s)^{-1} A_s^T$, we obtain $P_s = Q_s Q_s^T$. Once we written the projection matrix in this form, it is easy to show that for all $r \leq s$, we have $\bar{q}_r^T P_s = \bar{q}_r^T Q_s Q_s^T = \|\bar{q}_r\| \bar{e}_r^T Q_s^T = \bar{q}_r^T$. Taking the dot product of \bar{q}_{s+1} with \bar{q}_r , we obtain:

$$\bar{q}_r^T \bar{q}_{s+1} = \bar{q}_r^T \bar{a}_{s+1} - \bar{q}_r^T \bar{a}_{s+1} = 0$$

This proves the inductive assumption.

- 18.** Consider a $d \times d$ matrix A such that its right null space is identical to its column space. Show that d is even, and provide an example of such a matrix.

Let us say that the column space rank is k . Therefore, the row rank is k as well. Furthermore, the rank of the null space is k . Therefore, the sum of the row rank and null space is $2k$, which is d . An example of such a matrix is as follows:

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

- 19.** Show that the columns of the $n \times d$ matrix A are linearly independent if and only if $f(\bar{x}) = A\bar{x}$ is a one-to-one function.

Suppose that the columns of A are linearly independent and the function is not one-to-one. So we can find two different choices of \bar{x} equal to \bar{x}_1 and \bar{x}_2 , so that $f(\bar{x}_1) = f(\bar{x}_2)$. Therefore, it follows that $A(\bar{x}_1 - \bar{x}_2) = 0$, and it would imply that the columns of A are not linearly independent. A contradiction follows.

Now suppose that the function $f(\bar{x})$ is a one-to-one function, but the columns of A are not linearly independent. In such a case, a non-zero vector \bar{y} exists from the right null space of A , so that $A\bar{y} = \bar{0}$. Now consider any non-zero vector \bar{x} . It can be shown that $f(\bar{x} + \bar{y}) = A(\bar{x} + \bar{y}) = A\bar{x} + \bar{0} = f(\bar{x})$. In other words, we have $f(\bar{x} + \bar{y}) = f(\bar{x})$, which implies that $f(\cdot)$ is not a one-to-one function. A contradiction follows.

- 20.** Consider an $n \times n$ matrix A . Show that if the length of the vector $A\bar{x}$ is strictly less than that of the vector \bar{x} for all nonzero $\bar{x} \in \mathcal{R}^n$, then $(A - I)$ is invertible.

If $(A - I)$ is not invertible, its rank is strictly less than n and some nonzero vector \bar{v} lies in its null space. Therefore, $(A - I)\bar{v} = 0$. This would imply that $A\bar{v} = \bar{v}$, which is a contradiction to the pre-condition of the problem.

21. It is intuitively obvious that an $n \times n$ projection matrix P will always satisfy $\|P\bar{b}\| \leq \|\bar{b}\|$ for any $\bar{b} \in \mathcal{R}^n$, since it projects \bar{b} on a lower-dimensional hyperplane. Show algebraically that $\|P\bar{b}\| \leq \|\bar{b}\|$ for any $\bar{b} \in \mathcal{R}^n$.

Let us represent the projection matrix P in the form QQ^T for some $n \times d$ matrix Q , where $d \leq n$ and defines the dimensionality of the projection. In such a case, we have the following:

$$\|QQ^T\bar{b}\|^2 = \bar{b}^T \underbrace{Q^T Q}_I Q^T \bar{b} = \|Q^T \bar{b}\|^2. \text{ Note that } Q^T \bar{b} \text{ is a } d\text{-dimensional vector. We}$$

could add $(n-d)$ orthogonal columns to the $n \times d$ matrix Q to create an $n \times n$ matrix Q_{big} , which is a square orthogonal matrix and whose columns create a basis for all of \mathcal{R}^n rather than only the d -dimensional projection space. Note that $Q_{big}^T \bar{b}$ has the same first d coordinates as $Q^T \bar{b}$, and the former has the same length as \bar{b} because it is a square and orthogonal matrix. Therefore, the length $Q^T \bar{b}$ will be no larger than that of \bar{b} .

22. Let A be a 10×10 matrix. If A^2 has rank 6, find the minimum and maximum possible ranks of A . Give examples of both matrices.

By the Sylvester inequality the maximum rank of A is 8. This is because if the rank of A is a , the Sylvester inequality states that the minimum rank of A^2 is $2 * a - 10$. Therefore, a is at most 8, although it can be smaller for A^2 to meet the threshold. This bound is also tight, because examples exist where this is the case. Choosing A to be any strictly upper triangular matrix of rank 8 in which the bottom two rows are zeros, will result in A^2 with rank 6.

The minimum rank is 6. This is because A^2 can never have rank larger than that of A . This bound is also tight. Choosing A to be any projection matrix of rank 6 will result in $A^2 = A$, which also has rank 6.

23. Suppose that we have a system of equations $A\bar{x} = \bar{b}$ for some $n \times d$ matrix A . We multiply both sides of the above equation with a non-zero, $m \times n$ matrix B to obtain the new system $BA\bar{x} = B\bar{b}$. Provide an example to show that the solution sets to the two systems are not identical. How are the solution sets related in general? Provide one example of a sufficient condition on a rectangular matrix B under which they are identical.

An example is the case where we have an inconsistent system of equations $A\bar{x} = \bar{b}$. However, we choose $B = A^+$, which is the Moore-Penrose pseudoinverse of A . In such a case, the system becomes consistent after multiplying with B , and $\bar{x} = A^+ \bar{b}$ is a valid solution. Therefore, solutions often exist to the second system that are not present in the first system. However, any solution to the first system also satisfies the second system.

The solution set for the second system is always a superset of the first (and sometimes identical). The two solution sets are the same if B has linearly independent columns (sufficient condition).

The proof is as follows. If B has linearly independent columns, we can multiply both sides with the left-inverse of B to obtain the first system.

More detailed possibilities are as follows.

When the first system is consistent: The necessary and sufficient condition for identical solutions is that the column space of A and null space of B need to be

mutually exclusive. Note that there are two distinct solutions to the second equation only one of which belongs to the first, if and only if a vector $\bar{\alpha}$ (which is the difference of these solutions), exists for which $A\bar{\alpha} \neq 0$ and $BA\bar{\alpha} = 0$. This is possible if and only if the column space of A and null space of B are not mutually exclusive. Note that if B has linearly independent columns, its null space is empty.

When the first system is inconsistent: This means that \bar{b} cannot lie in the column space of A but must lie in the column space of BA . In other words, $\bar{b} \in \text{col}(BA) - \text{col}(A)$. Note that the column space of BA is a subset of the column space of B .

The overall conditions of this problem are rather messy, and therefore only a sufficient condition was asked.

24. Show that every $n \times n$ Householder reflection matrix can be expressed as $Q_1 Q_1^T - Q_2 Q_2^T$, where concatenating the columns of Q_1 and Q_2 creates an $n \times n$ orthogonal matrix, and Q_2 contains a single column. What is the nature of the linear transformation, when Q_2 contains more than one column?

Any Householder reflection matrix can be expressed as $(I - 2\bar{v}\bar{v}^T)$ for some column vector \bar{v} . Now find an orthogonal basis of \mathcal{R}^n in which \bar{v} is the first member, and Q_1 contains the remaining members. The matrix Q_2 is set to \bar{v} . The result is then true for this construction of $[Q_1, \bar{v}]$ because the identity matrix can be written as $I = Q_1 Q_1^T + \bar{v}\bar{v}^T$.

For the second part of the question, consider a matrix of the form $Q_1 Q_1^T - Q_2 Q_2^T$. This matrix can be written as $[Q_1, Q_2][Q_1, Q_2]^T - 2Q_2 Q_2^T = I - 2Q_2 Q_2^T$. Let the k columns of Q_2 be $\bar{v}_1 \dots \bar{v}_k$. The resulting matrix multiplication $Q_2 Q_2^T$ can be expanded in terms of outer products as $\sum_{i=1}^k \bar{v}_i \bar{v}_i^T$. Therefore, the resulting matrix is of the form $I - 2(\sum_{i=1}^k \bar{v}_i \bar{v}_i^T)$. Because of the orthogonality of $\bar{v}_1 \dots \bar{v}_k$, one can write this expression as $\prod_{i=1}^k (I - 2\bar{v}_i \bar{v}_i^T)$. This is simply a sequence of k Householder transforms.

25. Show that if B^k has the same rank as that of B^{k+1} for a particular value of $k \geq 1$, then B^k has the same rank as B^{k+r} for all $r \geq 1$.

We only need to show that B^{k+1} has the same rank as B^{k+2} , because we can apply the result repeatedly to prove that B^{k+2} has the same rank as $B^{k+3} \dots B^{k+r}$.

Suppose that the rank of B^{k+1} is not the same as that of B^{k+2} (and the latter must be smaller by the matrix product results discussed in the chapter). Then, there must exist a vector \bar{x} in the null space of B^{k+2} , which does not exist in the null space of B^{k+1} . Therefore, $B^{k+2}\bar{x} = \bar{0}$ and $B^{k+1}\bar{x} \neq \bar{0}$. However, this also means that we have a vector $\bar{y} = B\bar{x}$ for which we have $B^{k+1}\bar{y} = \bar{0}$ and $B^k\bar{y} \neq \bar{0}$. Therefore, we reach a contradiction to the statement of the assumption.

26. Show that if an $n \times n$ matrix B has rank $(n-1)$, and the matrix B^k has rank $(n-k)$, then each matrix B^r for r from 1 to k has rank $(n-r)$. Show how to construct a chain of vectors $\bar{v}_1 \dots \bar{v}_k$ so that $B\bar{v}_i = \bar{v}_{i-1}$ for $i > 1$, and $B\bar{v}_1 = \bar{0}$.

According to the Sylvester inequality, the rank of B^r can reduce by at most 1 because of multiplication with B . Therefore, the matrix B^k has rank that is at most $k-1$ less than that of B . However, since the rank of B^k is exactly $k-1$ less than that of B (according to the statement of the problem), it follows that the rank of B^r is exactly $(n-r)$.

The vector \bar{v}_k should be set so that $B^k \bar{v}_k = \bar{0}$ and $B^{k-1} \bar{v}_k \neq 0$. Such a vector always exists because the rank of B^k is strictly less than that of $(k-1)$. Subsequently, we set $\bar{v}_r = B^{k-r} \bar{v}_k$ for $r < k$.

- 27.** Suppose that $B^k \bar{v} = \bar{0}$ for a particular vector \bar{v} for some $k \geq 2$, and $B^r \bar{v} \neq \bar{0}$ for all $r < k$. Show that the vectors $\bar{v}, B\bar{v}, B^2\bar{v}, \dots, B^{k-1}\bar{v}$ must be linearly independent.

Let us define $\bar{v}_i = B^i \bar{v}$. Now consider the case where the vectors $\bar{v}_0 \dots \bar{v}_{k-1}$ are linearly dependent, and so we have $\sum_{i=0}^{k-1} \alpha_i \bar{v}_i = 0$. Multiplying both sides with B^{k-1} , we obtain $\alpha_0 B^{k-1} \bar{v}_0 = 0$. Therefore, $\alpha_0 = 0$. We again repeat this process by multiplying with B^{k-2} and prove that $\alpha_1 = 0$. We keep repeating until all coefficients are proven to be 0. In other words, the vectors are linearly independent.

- 28. Inverses with QR decomposition:** Suppose you perform QR decomposition of an invertible $d \times d$ matrix as $A = QR$. Show how you can use this decomposition relationship for finding the inverse of A by solving d different triangular systems of linear equations, each of which can be solved by backsubstitution. Show how to compute the left inverse and the right inverse of a (tall or fat) matrix with QR decomposition and back substitution.

Let $X = [\bar{x}_1 \dots \bar{x}_d]$ be the inverse of A . Then, we have $AX = I$, which we can write as follows:

$$\begin{aligned} A[\bar{x}_1 \dots \bar{x}_d] &= [\bar{e}_1 \dots \bar{e}_d] \\ QR[\bar{x}_1 \dots \bar{x}_d] &= [\bar{e}_1 \dots \bar{e}_d] \\ R[\bar{x}_1 \dots \bar{x}_d] &= [Q^T \bar{e}_1 \dots Q^T \bar{e}_d] \end{aligned}$$

Each triangular system of equations $R\bar{x}_i = Q^T \bar{e}_i$ is solved. This can be solved by back substitution.

For left inverse, we assume that the matrix A is of size $n \times d$ with $n > d$. The matrix A is decomposed as $A = QR$, where Q is of size $n \times d$ and R is of size $d \times d$. The left inverse is $(A^T A)^{-1} A^T$, which works out to $(R^T R)^{-1} R^T Q^T = R^{-1} Q^T$. Note that R^{-1} can be computed using backsubstitution as $RY = I_d$, where Y is the inverse of R .

For right inverse, we assume that the matrix A is of size $n \times d$ with $n < d$. However, here we decompose $A^T = QR$, where Q is of size $d \times n$ and R is of size $n \times n$. The right inverse is $A^T (AA^T)^{-1}$, which is $QR(R^T R)^{-1} = Q(R^T)^{-1}$. This is a similar situation to the left inverse except that we are computing in the inverse of R^T by backsubstitution.

- 29. Least-squares error by QR decomposition:** Let $A\bar{x} = \bar{b}$ be a system of equations in which the $n \times d$ matrix A has linearly independent columns. Suppose that you decompose $A = QR$, where Q is an $n \times d$ matrix with orthogonal columns and R is a $d \times d$ upper-triangular matrix. Show that the best-fit error (using the least-squares model) is given by $\|\bar{b}\|^2 - \|Q^T \bar{b}\|^2$. How would you find the least-squares error via QR decomposition in the case that A does not have linearly independent columns or rows?

The best fit of \bar{b} is simply its projection $QQ^T \bar{b}$ using the projection matrix QQ^T . Note that one can also derive this solution as $\bar{x} = R^{-1} Q^T \bar{b}$, and then multiply again with $A = QR$ to obtain $A\bar{x} = QQ^T \bar{b}$. Therefore, the best fit error is $\|A\bar{x} - \bar{b}\|^2 = (QQ^T \bar{b} - \bar{b})^T (QQ^T \bar{b} - \bar{b})$. On expanding with the distributive property, one obtains the desired result.

30. Consider a modified least-squares problem of minimizing $\|A\bar{x} - \bar{b}\|^2 + \bar{c}^T \bar{x}$, where A is an $n \times d$ matrix, \bar{x}, \bar{c} are d -dimensional vectors, and \bar{b} is an n -dimensional vector. Show that the problem can be reduced to the standard least-squares problem as long as \bar{c} lies in the row space of A . What happens when \bar{c} does not lie in the row space of A ?

When \bar{c} lies in the row space of A , we can rewrite \bar{c} as $A^T \bar{\alpha}$ for some column vector $\bar{\alpha}$. Therefore, the objective function becomes the following:

$$(A\bar{x} - \bar{b})^T (A\bar{x} - \bar{b}) + \bar{\alpha}^T A\bar{x} = (A\bar{x} - [\bar{b} - \bar{\alpha}/2])^T (A\bar{x} - [\bar{b} - \bar{\alpha}/2]) - \|\bar{\alpha}\|^2/4 + \bar{b}^T \bar{\alpha}$$

The last few terms in $\bar{\alpha}$ and \bar{b} are constants which do not affect the overall objective function. However, the vector \bar{b} has now been translated by $\bar{\alpha}/2$, which is the main difference to the original objective function.

If \bar{c} does not lie in the row space of A , the problem can have an unbounded minimum. An example is that of minimizing $(x_1 - 1)^2 + x_2$, where x_2 can be made as negative as we want.

31. **Right-inverse yields concise solution:** Let $\bar{x} = \bar{v}$ be any solution to the consistent system $A\bar{x} = \bar{b}$ with $n \times d$ matrix A containing linearly independent rows. Let $\bar{v}_r = A^T (AA^T)^{-1} \bar{b}$ be the solution given by the right inverse. Then, show the following:

$$\|\bar{v}\|^2 = \|\bar{v} - \bar{v}_r\|^2 + \|\bar{v}_r\|^2 + 2\bar{v}_r^T (\bar{v} - \bar{v}_r) \geq \|\bar{v}_r\|^2 + 2\bar{v}_r^T (\bar{v} - \bar{v}_r)$$

Now show that $\bar{v}_r^T (\bar{v} - \bar{v}_r) = 0$ and therefore $\|\bar{v}\|^2 \geq \|\bar{v}_r\|^2$.

The norm $\|\bar{v}\|^2$ can be written as $[\bar{v}_r + (\bar{v} - \bar{v}_r)]^T [\bar{v}_r + (\bar{v} - \bar{v}_r)]$, which can be expanded using the distributive property to obtain the first result. Note that the first inequality $\|\bar{v}\|^2 \geq \|\bar{v}_r\|^2 + 2\bar{v}_r^T (\bar{v} - \bar{v}_r)$ is obtained by dropping of the squared terms. Next, we substitute for \bar{v}_r in the expression $\bar{v}_r^T (\bar{v} - \bar{v}_r)$ to obtain the following:

$$\bar{v}_r^T (\bar{v} - \bar{v}_r) = [A^T (AA^T)^{-1} \bar{b}]^T (\bar{v} - \bar{v}_r) = \bar{b}^T (AA^T)^{-1} [\underbrace{A\bar{v}}_{\bar{b}} - \underbrace{A\bar{v}_r}_{\bar{b}}]$$

It is easy to see that the right-hand side evaluates to 0.

32. Show that any 2×2 Givens rotation matrix is a product of two Householder reflection matrices. Think geometrically before wading into the algebra. Now generalize the proof to $d \times d$ matrices.

Geometrically, if two mirrors facing each other (and intersecting at the origin) are aligned at an angle of $\theta/2$ or $90 - \theta/2$, then after two reflections in the mirrors, the angle of an object will change by a multiple of θ or $90 - \theta$. The exact angle depends on the order of reflection among the two mirrors, because it will not yield the same result. First, let us examine the Householder reflection matrix in which we want to compute $H_1 = (I - \bar{v}\bar{v}^T)$, where $\bar{v} = [\sin(\theta/2), \cos(\theta/2)]^T$. After computing the Householder reflection matrix, we obtain the following:

$$H_1 = \begin{bmatrix} 1 - 2\sin^2(\theta/2) & -2\cos(\theta/2)\sin(\theta/2) \\ -2\cos(\theta/2)\sin(\theta/2) & 1 - 2\cos^2(\theta/2) \end{bmatrix}$$

Using some trigonometric identities on the expressions for twice the angles, we obtain the following Householder reflection matrix:

$$H_1 = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ -\sin(\theta) & -\cos(\theta) \end{bmatrix}$$

This is already very close to a Givens matrix except that the handedness of the data is wrong. Therefore we choose the following matrix:

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

It is easy to see that H_2H_1 is a Givens rotation at counter-clockwise angle θ for column vectors:

$$H_2H_1 = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

The result can be directly generalized to $d \times d$ Givens rotations, since Givens rotations are performed in 2 dimensions, and we only need to select out the relevant two dimensions in both the rotations and reflections. This is achieved by selecting the four non-zero entries in the Givens matrix by using the relevant two dimensions. Further, one can set all except two entries in the vector \bar{v} to 0. The entries that are set to 0 correspond to irrelevant dimensions. The other two entries are $\sin(\theta/2)$ and $\cos(\theta/2)$. The matrix H_2 is also set by setting one dimension of the identity matrix to -1 so that a Givens rotation is created.

- 33.** *Show that if two tall matrices of full rank have the same column space, then they have the same projection matrix.*

If two matrices A and C have the same column space, then a non-singular matrix B exists so that $AB = C$. Now show that the projection matrix of AB and A are the same. In fact, this proof is explicitly shown in the text.

- 34.** *Construct 4×3 matrices A and B of rank 2 that are not multiples of one another, but with the same four fundamental subspaces of linear algebra.*

We should select any 2-dimensional subspace of \mathcal{R}^4 and put two possible basis sets in the columns of the 4×2 matrices U_1 and U_2 . Similarly, we should select any 2-dimensional subspace of \mathcal{R}^3 , and create two possible basis sets in the rows of V_1 and V_2 . Then, the matrices U_1V_1 and U_2V_2 will have the same four fundamental subspaces of linear algebra.

- 35.** *Show that any Householder reflection matrix $(I - 2\bar{v}\bar{v}^T)$ can be expressed as follows:*

$$(I - 2\bar{v}\bar{v}^T) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}$$

Relate \bar{v} to θ geometrically.

Select $\bar{v} = [-\sin(\theta/2), \cos(\theta/2)]^T$, and show that $(I - 2\bar{v}\bar{v}^T)$ turns out to be the matrix in the statement of the problem by using standard trigonometric identities. Note that the reflection is performed on a line making a counter-clockwise angle of $\theta/2$ about the X -axis.

- 36.** *Show how any vector $\bar{v} \in \mathcal{R}^n$ can be transformed to $\bar{w} \in \mathcal{R}^n$ as $\bar{w} = cH\bar{v}$, where c is a scalar and H is an $n \times n$ Householder reflection matrix.*

First, the vectors \bar{v} and \bar{w} are scaled to unit norm to create \bar{v}_1 and \bar{w}_1 , respectively. Then, the Householder reflection matrix is created using $(\bar{v}_1 - \bar{w}_1)$ as $(I - 2((\bar{v}_1 - \bar{w}_1)(\bar{v}_1 - \bar{w}_1)^T))$. The scalar c simply adjusts for the varying length of the two vectors, and is chosen to be $||\bar{w}||/||\bar{v}||$.

37. A block upper-triangular matrix is a generalization of a block diagonal matrix that allows nonzero entries above the square, diagonal blocks. Consider a block upper-triangular matrix with invertible diagonal blocks. Make an argument why such a matrix is row equivalent to an invertible block diagonal matrix. Generalize the backsubstitution method to solving linear equations of the form $A\bar{x} = \bar{b}$ when A is block upper-triangular. You may assume that the diagonal blocks are easily invertible.

One should reduce the rows from bottom to top. The goal is to reduce all non-diagonal elements to zero. The lowest block of diagonal elements is left alone. The non-zero elements of the block of diagonal elements can be reduced it by using the fact that a basis exists for that subspace in the diagonal block just below it. Similarly we can reduce two non-diagonal blocks just above it, and continue the process until all diagonal blocks are reduced.

One can generalize backsubstitution in a relatively simple manner by first solving for the variables in the lowest block by an inversion. Next, we can substitute this variables and solve for the next higher block. The process is repeated to solve for all the variables.

38. If P is a projection matrix, show that $(P + \lambda I)$ is invertible for any $\lambda > 0$.

This can be easily shown by expressing P as QQ^T . The idea is to show that $\bar{x}^T(P + \lambda I)\bar{x} = \|Q^T\bar{x}\|^2 + \lambda\|\bar{x}\|^2 > 0$ for all \bar{x} and therefore $(P + \lambda I)\bar{x} \neq \bar{0}$. This means that $P + \lambda I$ has an empty null space.

39. If R is a Householder reflection matrix, show that $(R + I)$ is always singular, and that $(R + \lambda I)$ is invertible for any $\lambda \notin \{1, -1\}$.

Note that $(R + I)\bar{v}$ is zero, when \bar{v} is the unit vector used for constructing the Householder reflection matrix.

Now we will provide non-singularity of $R + \lambda I$. Decompose the vector \bar{x} to $c\bar{v} + \bar{v}_2$, where $c\bar{v}$ is parallel to \bar{v} and \bar{v}_2 is perpendicular to \bar{v} . The value of $(R + \lambda I)\bar{x}$ can be shown to be $c(\lambda - 1)\bar{v} + \bar{v}_2(\lambda + 1)$. Note that this vector is always non-zero when λ is not equal to either $+1$ or -1 .

40. **Length-preserving transforms are orthogonal:** We already know that if A is an $n \times n$ orthogonal matrix, then $\|A\bar{x}\| = \|\bar{x}\|$ for all $\bar{x} \in \mathcal{R}^n$. Prove the converse of this result that if $\|A\bar{x}\| = \|\bar{x}\|$ for all $\bar{x} \in \mathcal{R}^n$, then A is orthogonal.

Let the columns of A be $\bar{a}_1 \dots \bar{a}_n$. By the pre-condition of the lemma, we know that the length of $A\bar{e}_i$ is the same as the length of \bar{e}_i where \bar{e}_i has a single 1 in the i th position and 0 otherwise. In other words, the length of $A\bar{e}_i$ is 1, or in other words, the length of \bar{a}_i is 1. Furthermore, the squared length of $\bar{a}_i^T A$ is the same as that of \bar{a}_i . This means the following:

$$\underbrace{\|\bar{a}_i\|^4}_1 + \sum_{j:j \neq i} (\bar{a}_i \cdot \bar{a}_j)^2 = \underbrace{\|\bar{a}_i\|^2}_1$$

Therefore, we obtain the following:

$$\sum_{j:j \neq i} (\bar{a}_i \cdot \bar{a}_j)^2 = 0$$

This is possible only if the dot product of \bar{a}_i with each vector \bar{a}_j is 0. In other words, each column of A is orthogonal to every other column. We have already proven that each column has unit length. Therefore, A is an orthogonal matrix.

41. Let A be a square $n \times n$ matrix so that $(A + I)$ has rank $(n - 2)$. Let $f(x)$ be the polynomial $f(x) = x^3 + x^2 + x + 1$. Show that $f(A)$ has rank at most $(n - 2)$. Furthermore, show that $f(A)$ has rank exactly $(n - 2)$ if A is symmetric.

The polynomial function $f(A)$ can be represented as follows:

$$f(A) = (A + I)(A^2 + I)$$

The rank of $f(A)$ is at most that of $(A + I)$ (by the matrix multiplication rule of ranks), which is $(n - 2)$. Furthermore, when A is symmetric the other factor $(I + A^2) = (I + A^T A)$ is non-singular. This is because for any non-zero vector \bar{x} we have $\bar{x}^T (I + A^T A) \bar{x} = \|\bar{x}\|^2 + \|A\bar{x}\|^2 > 0$. This means that $(I + A^T A)\bar{x}$ is non-zero for any \bar{x} , and therefore the matrix is non-singular. Since the factor $(I + A^2)$ is non-singular, the rank of $f(A) = (I + A)(A + A^2)$ is exactly equal to that of $(I + A)$.

42. Suppose that a $d \times d$ matrix A exists along with d vectors $\bar{x}_1 \dots \bar{x}_d$ so that $\bar{x}_i^T A \bar{x}_j$ is zero if and only if $i \neq j$. Show that the vectors $\bar{x}_1 \dots \bar{x}_d$ are linearly independent. Note that A need not be symmetric.

Suppose that the vectors are linearly dependent. Then scalars $\beta_1 \dots \beta_d$ exist so that $\sum_{i=1}^d \beta_i \bar{x}_i = 0$. Left-multiplying both sides with $\bar{x}_1^T A$, we obtain $\beta_1 \underbrace{(\bar{x}_1^T A \bar{x}_1)}_{\neq 0} = 0$.

This means that β_1 is zero. We can use a similar approach to prove each β_i to be 0. Therefore, the vectors are linearly independent.

43. Suppose that a $d \times d$ symmetric matrix S exists along with d vectors $\bar{x}_1 \dots \bar{x}_d$ so that $\bar{x}_i^T S \bar{x}_j$ is zero when $i \neq j$ and positive when $i = j$. Show that $\langle \bar{x}, \bar{y} \rangle = \bar{x}^T S \bar{y}$ is a valid inner product over all $\bar{x}, \bar{y} \in \mathcal{R}^d$.

The main point is to show the axioms for inner products. The commutative axiom follows from the symmetric nature of S . The multiplicative and distributive axioms follow from the simple algebra of $\bar{x}^T S \bar{y}$. For the positive definite axiom, we need to express an arbitrary vector $\bar{x} = \sum_{i=1}^d \alpha_i \bar{x}_i$. This is possible only because $\bar{x}_1 \dots \bar{x}_d$ are linearly independent and span all of \mathcal{R}^d . Then, one can show that $\bar{x}^T S \bar{x} = \sum_{i=1}^d \alpha_i^2 \bar{x}_i^T S \bar{x}_i + \sum_i \sum_{j \neq i} \alpha_i \alpha_j \bar{x}_i^T S \bar{x}_j$. The first term is positive and the second term is zero. The result follows.

44. **Cauchy-Schwarz and triangle inequality for general inner products:** Let \bar{u} and \bar{v} be two vectors for which $\langle \bar{u}, \bar{u} \rangle = \langle \bar{v}, \bar{v} \rangle = 1$. Show using only the inner-product axioms that $|\langle \bar{u}, \bar{v} \rangle| \leq 1$. Now show the more general Cauchy-Schwarz inequality by defining \bar{u} and \bar{v} appropriately in terms of \bar{x} and \bar{y} :

$$|\langle \bar{x}, \bar{y} \rangle| \leq \sqrt{\langle \bar{x}, \bar{x} \rangle \langle \bar{y}, \bar{y} \rangle}$$

Now use this result to show the triangle inequality for the triangle formed by \bar{x} , \bar{y} , and the origin:

$$\sqrt{\langle \bar{x}, \bar{x} \rangle} + \sqrt{\langle \bar{y}, \bar{y} \rangle} \geq \sqrt{\langle \bar{x} - \bar{y}, \bar{x} - \bar{y} \rangle}$$

The first part can be shown by using the fact that both $\langle \bar{u} + \bar{v}, \bar{u} + \bar{v} \rangle$ and $\langle \bar{u} - \bar{v}, \bar{u} - \bar{v} \rangle$ are non-negative by the positive definiteness axiom. Note that these values *can* be zero, if $\bar{u} = \pm \bar{v}$. Therefore, we are not using strict inequality for the positive definiteness

axiom. Now, distributing the expressions for $\langle \bar{u} + \bar{v}, \bar{u} + \bar{v} \rangle$ and $\langle \bar{u} - \bar{v}, \bar{u} - \bar{v} \rangle$, we obtain $2 + 2\langle \bar{u}, \bar{v} \rangle$ and $2 - 2\langle \bar{u}, \bar{v} \rangle$, which can both be nonnegative only when we have $|\langle \bar{u}, \bar{v} \rangle| \leq 1$. Next, for arbitrary vectors \bar{x} and \bar{y} , we define \bar{u} and \bar{v} as follows:

$$\begin{aligned}\bar{u} &= \bar{x} / \sqrt{\langle \bar{x}, \bar{x} \rangle} \\ \bar{v} &= \bar{y} / \sqrt{\langle \bar{y}, \bar{y} \rangle}\end{aligned}$$

These two vectors can be verified to have unit norm, and therefore, we have $|\langle \bar{u}, \bar{v} \rangle| \leq 1$. By substituting the values of \bar{u} and \bar{v} in terms of \bar{x} and \bar{y} , we obtain the Cauchy-Schwarz inequality.

Starting with *twice* the Cauchy Schwarz inequality, we can add the sum of squared norms of \bar{x} and \bar{y} from both sides to obtain the following:

$$2|\langle \bar{x}, \bar{y} \rangle| + \langle \bar{x}, \bar{x} \rangle + \langle \bar{y}, \bar{y} \rangle \leq 2\sqrt{\langle \bar{x}, \bar{x} \rangle \langle \bar{y}, \bar{y} \rangle} + \langle \bar{x}, \bar{x} \rangle + \langle \bar{y}, \bar{y} \rangle$$

We can remove the modulus without affecting the inequality while using $-\langle \bar{x}, \bar{y} \rangle$ in lieu of $|\langle \bar{x}, \bar{y} \rangle|$. Therefore, we have the following:

$$-2\langle \bar{x}, \bar{y} \rangle + \langle \bar{x}, \bar{x} \rangle + \langle \bar{y}, \bar{y} \rangle \leq 2\sqrt{\langle \bar{x}, \bar{x} \rangle \langle \bar{y}, \bar{y} \rangle} + \langle \bar{x}, \bar{x} \rangle + \langle \bar{y}, \bar{y} \rangle$$

Using the distributive axiom of inner products, we have shown the following:

$$\begin{aligned}\langle \bar{x} - \bar{y}, \bar{x} - \bar{y} \rangle &\leq 2\sqrt{\langle \bar{x}, \bar{x} \rangle \langle \bar{y}, \bar{y} \rangle} + \langle \bar{x}, \bar{x} \rangle + \langle \bar{y}, \bar{y} \rangle \\ &\leq (\sqrt{\langle \bar{x}, \bar{x} \rangle} + \sqrt{\langle \bar{y}, \bar{y} \rangle})^2\end{aligned}$$

Taking the square-root of both sides, we obtain the desired result.

45. If the matrix computed by the polynomial function $f(A) = \sum_{i=0}^d c_i A^i$ has rank strictly greater than that of A , is there anything you can say about the coefficients $c_0 \dots c_d$?

The value of c_0 must be non-zero, or else A becomes a factor of the polynomial expression. In the latter case, the rank is bounded above by that of A .

46. Let S be a symmetric matrix and $g(S) = S^3 - S^2 + S$. Without using the results of the next chapter, show that $g(S)$ has the same rank as S .

We write $g(S)$ as the product of S and $f(S) = S^2 - S + I$. The function $f(S)$ can be written as $f(S) = (S - I/2)^2 + 3I/4 = (S - I/2)^T(S - I/2) + 3I/4$. Therefore, one can show that $\bar{x}^T f(S) \bar{x} = \|(S - I/2)\bar{x}\|^2 + 3\|\bar{x}\|^2/4$. This value is always positive for non-zero \bar{x} . This also means that $f(S)\bar{x}$ is always non-zero for non-zero \bar{x} . In other words, $f(S)$ has an empty null space and is non-singular. The product of S and a non-singular matrix has the same rank as S .

47. Let A be an $n \times m$ matrix and B be a $k \times d$ matrix. Show that the column space of AXB is always a subspace of the column space of A , and the row space of AXB is a subspace of the row space of B for any $m \times k$ matrix X .

Note that each column of AXB is a linear combination of the columns of A , by setting $Y = XB$ and using the columns of Y as the linear combination coefficients. The same holds true for the row space.

48. Suppose that A is an $n \times m$ matrix and B is a $k \times d$ matrix, both of full rectangular rank. You want to find the $m \times k$ matrix X so that $AXB = C$, where C is a known $n \times d$ matrix. What should the shapes of each of A and B be (i.e., tall, square, or wide) for the system of equations to be guaranteed to have at least one consistent solution? Derive an expression for one solution, X , in terms of A , B , and C in this case. When is this solution unique?

A should be wide and B should be tall. This ensures that the column space of A spans all of \mathcal{R}^n and the row space of B spans all of \mathcal{R}^d . Therefore, the column space of C is a subspace of A and the row space of C is a subspace of B . Therefore, by using the construction of the previous exercise, at least one solution X exists to this system of equations. In fact, we can even get a closed-form solution because of the full rank of the rectangular matrices in this particular case. A possible solution is $X = A^T(AA^T)^{-1}C(B^TB)^{-1}B^T$. Note that we are left-multiplying C with the right inverse of A , and right multiplying with the left-inverse of B . It is easy to verify that $AXB = C$ by plugging in the expression for X . The solution will not be unique unless A and B are both square. In such a case, the solution simplifies to $X = A^{-1}CB^{-1}$.

49. Suppose that A is an $n \times m$ matrix and B is a $k \times d$ matrix, both of full rectangular rank. A is tall and B is wide. The system of equations is inconsistent. You want to find the best-fit $m \times k$ matrix X so that $\|C - AXB\|_F^2$ is as small as possible, where C is a known $n \times d$ matrix. So you set $Y \approx XB$, and first find the best-fit solution to $\|C - AY\|_F^2$ and then find the best-fit solution to $\|Y - XB\|_F^2$. You use the normal equations to derive closed-form expressions for X and Y . Show that the closed-form solution for X and the best-fit C' to C are as follows:

$$X = \underbrace{(A^T A)^{-1} A^T}_{\text{Left Inverse}} C \underbrace{B^T (B B^T)^{-1}}_{\text{Right inverse}}, \quad C' = \underbrace{A (A^T A)^{-1} A^T}_{\text{Project columns}} C \underbrace{B^T (B B^T)^{-1} B}_{\text{Project rows}}$$

Note that the problem of optimizing $\|C - AY\|_F^2$ can be decomposed into vanilla least-squares problems $\|\bar{c}_i - A\bar{y}_i\|$ using the columns of C and Y , respectively. Therefore, we get $\bar{y}_i = (A^T A)^{-1} A^T \bar{c}_i$ by using the normal equations. Note that $A^T A$ is invertible because it is tall and of full rank. This yields $Y = (A^T A)^{-1} A^T C$. Next, we decompose the problem of minimizing $\|Y - XB\|_F^2$ using the rows of Y and B . Therefore, we get $X = C B^T (B B^T)^{-1}$. Combining the two, we get the desired result. The best-fit C' is AXB .

50. **Challenge Problem:** Let A be an $n \times m$ matrix and B be a $k \times d$ matrix. You want to find the $m \times k$ matrix X so that $C = AXB$, where C is a known $n \times d$ matrix. Nothing is known about the linear independence of rows or columns of A , B , and C . Propose a variation of the Gaussian elimination method to solve the system of equations $C = AXB$. How can you recognize inconsistent systems of equations or systems with an infinite number of solutions?

This can be achieved by using row reduction of A and column reduction of B . Note that the echelon form of A is a row-centric echelon form, and the echelon form of B is a column-centric echelon form. The transpose of a row-centric echelon form is a column-centric echelon form. Every time a row operation is performed on A it is performed on C , and every time a column operation is performed on B , it is performed on C . At the end of the process, we will obtain the following system:

$$A'XB' = C'$$

Here A' is in row echelon form and B' is in column echelon form. The system is inconsistent if and only if (i) the i th row in A' is zero, but the i th row in C' is not zero, or (ii) the j th column in B' is zero, but the j th column in C' is non-zero.

Assume that the system is consistent with matching zero rows and columns. We first get rid of the matching zero rows and columns from A' , B' and C' . The system has an infinite number of solutions if there are free columns in A' or if there are free rows in B' . One can write the reduced form of the equations as follows:

$$[U_A F_A] \begin{bmatrix} X_{11} & X_{12}^F \\ X_{21}^F & X_{22}^F \end{bmatrix} \begin{bmatrix} L_B \\ F_B \end{bmatrix} = C'$$

Here, U_A is a square, invertible upper-triangular matrix, F_A are the free columns of A , L_B is a square, invertible lower triangular matrix, and F_B are the free columns of B . Similarly, the matrix X has been partitioned into one block X_{11} of non-free variables, and three blocks of free variables, each of which is super-scripted by F . We can expand the above equation as follows:

$$\begin{aligned} U_A X_{11} L_B + U_A X_{12} F_B + F_A X_{21} L_B + F_A X_{22} F_B &= C' \\ U_A X_{11} L_B &= C' - U_A X_{12} F_B - F_A X_{21} L_B - F_A X_{22} F_B \end{aligned}$$

The free variables blocks for X can be set to any values we want to matrices Λ , Σ and Γ :

$$U_A X_{11} L_B = C' - U_A \Lambda F_B - F_A \Sigma L_B - F_A \Gamma F_B$$

Then, X_{11} can be found using two application of back substitution. Alternatively, if a closed form is desired, we can invert the upper and lower triangular matrices U_A and L_B to obtain the following:

$$X_{11} = U_A^{-1} \{C' - U_A \Lambda F_B - F_A \Sigma L_B - F_A \Gamma F_B\} L_B^{-1}$$

- 51.** Use the limit-based definition of the Moore-Penrose pseudoinverse to show that $A^T A A^+ = A^T$ and $B^+ B B^T = B^T$.

We only show the first of the two results, because the second is very similar. We have the following:

$$\begin{aligned} A^T A A^+ &= \lim_{\lambda \rightarrow 0^+} A^T A (A^T A + \lambda I)^{-1} A^T \\ &= \lim_{\lambda \rightarrow 0^+} (A^T A + \lambda I - \lambda I) (A^T A + \lambda I)^{-1} A^T \\ &= \lim_{\lambda \rightarrow 0^+} (A^T A + \lambda I) (A^T A + \lambda I)^{-1} A^T - \lambda (A^T A + \lambda I)^{-1} A^T \\ &= A^T - \lim_{\lambda \rightarrow 0^+} \lambda (A^T A + \lambda I)^{-1} A^T \\ &= A^T - [\lim_{\lambda \rightarrow 0^+} \lambda] \underbrace{[\lim_{\lambda \rightarrow 0^+} (A^T A + \lambda I)^{-1} A^T]}_{A^+} \\ &= A^T - A^+ \lim_{\lambda \rightarrow 0^+} \lambda = A^T \end{aligned}$$

Note that this proof is very simple if we are allowed to use singular value decomposition and substitute $A = Q \Sigma P^T$ and $A^+ = P \Sigma^+ Q^T$. However, this method is not used, since it is not covered in this chapter. Nevertheless, it would be a useful exercise to prove this using SVD, for those who are familiar with the technique.

- 52.** We know that the best-fit solution to $A\bar{x} = \bar{b}$ is given by $\bar{x}^* = A^+\bar{b}$. Therefore, we have $A\bar{x}^* = AA^+\bar{b}$. Show that the matrix AA^+ is both symmetric and idempotent (which is an alternative definition of a projection matrix). What type of projection does AA^+ perform here?

The matrix AA^+ is idempotent because we have $(AA^+)^2 = (AA^+A)A^+$. Now, one can use the limit-based definition of the Moore-Penrose pseudoinverse in order to show that $AA^+A = A$. The proof of this is very similar to that of the previous exercise. Therefore, the idempotent property is satisfied.

In order to show the symmetric property, note that we have:

$$(AA^+)^T = \lim_{\lambda \rightarrow 0^+} [AA^T(AA^T + \lambda I)^{-1}]^T = \lim_{\lambda \rightarrow 0^+} [(AA^T + \lambda I)^{-1}]^T AA^T$$

Now note that $(AA^T + \lambda I)$ is symmetric, and the inverse of a symmetric matrix is a symmetric matrix as well (see Exercise 22 of Chapter 1). Therefore, applying the transposition to this inverse does not matter. Therefore, we can remove the transposition condition on this inverse to simplify the above as follows:

$$(AA^+)^T = \lim_{\lambda \rightarrow 0^+} (AA^T + \lambda I)^{-1} AA^T$$

Now, based on Exercise 41 of Chapter 1, the matrix $(AA^T + \lambda I)^{-1}$ and AA^T commute, because they are polynomial functions of the same matrix AA^T . Therefore, we can further simplify the above to the following:

$$(AA^+)^T = \lim_{\lambda \rightarrow 0^+} AA^T(AA^T + \lambda I)^{-1} = A \lim_{\lambda \rightarrow 0^+} A^T(AA^T + \lambda I)^{-1} = AA^+$$

Therefore, AA^+ is symmetric as well. This makes AA^+ a projection matrix. Note that this proof is very simple if we are allowed to use singular value decomposition and substitute $A = Q\Sigma P^T$ and $A^+ = P\Sigma^+Q^T$.

This projection matrix projects the vector \bar{b} into the column space of A , and it works even when the columns of A are not linearly independent. The basic idea is that the best fit of \bar{b} is always its projection into the column space of A irrespective of whether or not A is of full rank. The Moore-Penrose pseudoinverse provides a way to compute the projection matrix even for matrices that do not have linearly independent columns.

Chapter 3

Diagonalizable Matrices and Eigenvectors

1. Any $d \times d$ matrix A can be decomposed into $O(d^2)$ Givens rotations and at most one elementary reflection. Discuss how the sign of the determinant of A determines whether or not a reflection is needed.

The sign of the determinant flips when a reflection is needed.

2. Any $d \times d$ matrix A can be decomposed into at most $O(d)$ Householder reflections. Discuss the effect of the sign of the determinant on the number of Householder reflections.

An odd number of Householder reflections will be performed when the sign of the determinant flips.

3. Show that if a matrix A satisfies $A^2 = 4I$, then the eigenvalues of A include at least one of the values 2 and -2 .

It is easy to show that $(A + 2I)(A - 2I) = 0$. Then, for any vector \bar{x} , we know that $(A + 2I)(A - 2I)\bar{x} = 0$. One of the two cases must hold:

$(A - 2I)\bar{x} = 0$: In this case, it is clear that the matrix A has eigenvalue 2.

$(A - 2I)\bar{x} \neq 0$: In this case, we can set this nonzero vector to \bar{y} . It is evident that $(A + 2I)\bar{y} = 0$. Therefore, A has eigenvalue of -2 .

4. You are told that a 4×4 symmetric matrix has eigenvalues 4, 3, 2, and 2. You are given the values of eigenvectors belonging to the eigenvalues 4 and 3. Provide a procedure to reconstruct the entire matrix.

Since the matrix is symmetric, its eigenvectors must be orthogonal. Furthermore, the eigenspace for eigenvalue 2 has dimensionality 2. Therefore, one can pick any pair of orthogonal vectors that are also orthogonal to the known eigenvectors as the eigenvectors of eigenvalue 2. These four eigenvectors can then be used to construct the eigenvector matrix V . The reconstructed matrix is then $V\Delta V^T$.

5. Suppose that A is a square $d \times d$ matrix. The matrix A' is obtained by multiplying the i th row of A with γ_i and dividing the i th column of A with γ_i for each i . Discuss how the eigenvectors of A are related to those of A' .

The i th component of the eigenvector will get multiplied with γ_i . This is because we have $A' = \Delta A \Delta^{-1}$, and the two matrices are similar. Here, Δ is a diagonal matrix in which the i th entry is γ_i . It can be shown that if \bar{x} is an eigenvector of A , then $\Delta \bar{x}$ is an eigenvector of A' .

6. For a 4×4 matrix A with the following list of eigenvalues obtained from the characteristic polynomial, state in each case whether the matrix is guaranteed to be diagonalizable, invertible, both, or neither: (a) $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{1, 3, 4, 9\}$ (b) $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{1, 3, 3, 9\}$ (c) $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{0, 3, 4, 9\}$ (d) $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{0, 3, 3, 9\}$ (e) $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{0, 0, 4, 9\}$. —

(a) both, (b) invertible (c) diagonalizable (d) neither (e) neither

In each case, distinct eigenvalues are necessary to guarantee diagonalizability. Nonzero eigenvalues ensure invertibility.

7. Show that any real-valued matrix of odd dimension must have at least one real eigenvalue. Show the related fact that the determinant of a real-valued matrix without any real eigenvalues is always positive. Furthermore, show that a real-valued matrix of even dimension with a negative determinant must have at least two distinct real-valued eigenvalues.

Note that the expression $\det(A - \lambda I)$ is a polynomial of odd degree, which has at least one real root. Furthermore, for the case when the determinant of a real matrix does not have real eigenvalues, the degree must be even. Furthermore, the value of $\det(A - \lambda I)$ is a polynomial of odd degree in which the highest (even) degree of the polynomial has a positive coefficient. Therefore, for large absolute values of the polynomial, the value of $\det(A - \lambda I)$ is positive. Since the polynomial has no real roots, the entire polynomial is positive for all values of λ including at $\lambda = 0$. In other words, $\det(A)$ is positive. In the event that the determinant is negative, the polynomial is negative at $\lambda = 0$. Therefore, it must have at least one positive root and one negative root in order for $\det(A - \lambda I)$ to switch to positive values at extreme values of λ .

8. Consider the Jordan normal form $A = VUV^{-1}$. Show that the upper triangular matrix U is in block diagonal form, where smaller upper-triangular matrices $U_1 \dots U_r$ are arranged along the diagonal of U , and other entries are zeros. What is the effect of applying a polynomial function $f(U)$ on the individual blocks $U_1 \dots U_r$? Use this fact to provide a general proof of the Cayley-Hamilton theorem.

Applying a polynomial on a block diagonal matrix is equivalent to applying the polynomial to each block. Now note that for a block of size $r \times r$ with eigenvalue λ , the matrix $(A - \lambda I)^r$ is a factor of the characteristic polynomial. The matrix $(A - \lambda I)^r$ is strictly triangular and nil potent. Therefore, its r th power is the zero matrix. In other words, after applying the characteristic polynomial to each block, it will be set to 0. The proof follows.

9. Provide an example of a defective matrix whose square is diagonalizable.

Consider a defective matrix A of size 2×2 , which is strictly upper triangular with a value of 1 as its only non-zero entry in the upper corner. This matrix has eigenvalue 0 with multiplicity of 2, and is not diagonalizable. Its square is the zero matrix, which is diagonalizable. In fact, the matrix A is already (trivially) in Jordan normal form, and one can use any invertible as V , so that $B = VAV^{-1}$. Squaring this matrix results in the 2×2 matrix of zeros. In general, if the only generalized eigenvectors corresponds to

a Jordan block with zero eigenvalues, this becomes a possibility. Therefore, non-trivial cases of this scenario can also be constructed by creating a single repeated eigenvalue of zero and an eigenvalue of 1. Consider any matrix A of the following form with arbitrary invertible matrix V :

$$A = V \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} V^{-1}$$

This matrix is in Jordan normal form and is not diagonalizable. However, its square is of the following diagonalizable form:

$$A^2 = V \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} V^{-1}$$

- 10.** Let A and B be $d \times d$ matrices. Show that the matrix $AB - BA$ can never be positive semi-definite unless it is the zero matrix.

The trace of $AB - BA$ is 0. This means that the sum of the eigenvalues is 0. The sum of the eigenvalues of a positive semi-definite matrix is 0 if and only if all eigenvalues are 0. This is true if and only if the matrix contains only zero entries.

- 11.** Can the square of a matrix that does not have real eigenvalues be diagonalizable with real eigenvalues? If no, provide a proof. If yes, provide an example.

Yes, it is possible. A matrix corresponding to a 90° rotation is not diagonalizable with real eigenvalues. However, the square of this matrix is a 180° rotation, which is the negative of the identity matrix. This matrix is already diagonal.

- 12.** Let A and B be two diagonalizable matrices. Then, A and B are simultaneously diagonalizable if and only if $AB = BA$. You may assume for simplicity that algebraic multiplicity of each eigenvalue is 1 (although the result is true in general). Further show that if the matrices A , B , and AB are all symmetric, show that the matrices A and B must be simultaneously diagonalizable.

If the two matrices are simultaneously diagonalizable as $A = V\Delta_1V^{-1}$ and $B = V\Delta_2V^{-1}$, we have:

$$AB = V\Delta_1\Delta_2V^{-1} = V\Delta_2\Delta_1V^{-1} = BA$$

In other words, the matrices commute.

Now consider the case where the two matrices commute. It suffices to show that every eigenvector of A is an eigenvector of B , and vice versa. If \bar{x} is an eigenvector of B with eigenvalue λ , we have:

$$B[A\bar{x}] = A[B\bar{x}] = \lambda A\bar{x}$$

In other words, $A\bar{x}$ is an eigenvector of B with the *same eigenvalue*. To ensure an algebraic multiplicity of 1 for eigenvectors of B with eigenvalue λ , $A\bar{x}$ must be a scalar multiple of \bar{x} . In other words, \bar{x} is an eigenvector of A as well.

According to the exercise mentioned in the problem hint, we must have $AB = BA$. Then, according to the above result, since A and B are symmetric and $AB = BA$, the matrices A and B are simultaneously diagonalizable.

13. Suppose that the $d \times d$ matrix S is symmetric, positive semi-definite matrix, and the matrix D is of size $n \times d$. Show that DSD^T must also be a symmetric, positive semi-definite matrix.

It is easy to show that DSD^T is symmetric because $(DSD^T)^T = DS(D^T)^T = DS$. For positive semi-definiteness, note that for any d -dimensional vector \bar{x} , we have $\bar{x}^T DSD^T \bar{x} = \bar{y}^T S \bar{y}$ by setting $D^T \bar{x} = \bar{y}$. However, since S is positive semi-definite, we must have $\bar{y}^T S \bar{y} \geq 0$. The result follows.

14. Let S be a positive semi-definite matrix, which can therefore be expressed in Gram matrix form as $S = B^T B$. Use this fact to show that the diagonal entries of a positive semi-definite matrix can never be negative, and further show that positive definite matrices cannot even have zero entries.

Each diagonal entry of S is of the form $\bar{b}_i^T \bar{b}_i = \|\bar{b}_i\|^2$, where \bar{b}_i is the i th column of B . This value can never be negative. Furthermore, this value can be 0 only when \bar{b}_i is 0 and B is singular. However, B cannot be singular for a positive definite matrix.

15. Show that if a matrix P satisfies $P^2 = P$, then all its eigenvalues must be 1 or 0.

Let (\bar{x}, λ) be an eigenvector-eigenvalue pair of P . Since we have $P^2 - P = 0$, it follows that $(P^2 - P)\bar{x} = 0$. In other words, we have $\lambda^2 \bar{x} - \lambda \bar{x} = 0$. This is possible only when we have $\lambda^2 - \lambda = 0$. In other words, we have $\lambda \in \{0, 1\}$.

16. Show that a matrix A is always similar to its transpose A^T . A hint for solving this problem is that a similar family is uniquely defined by its Jordan normal form. Therefore, one can show that the upper triangular matrix in Jordan normal form is similar to its transpose.

The Jordan normal form upper-triangular matrix can be shown to be similar to its transpose, because reversing the order of the rows and columns of the Jordan blocks yields the transpose of the Jordan upper-triangular matrix. Therefore, the two matrices are similar by the use of a permutation matrix as the basis change. Since the Jordan normal form upper-triangular matrix and its transpose are similar with the use of a permutation matrix as the basis change, it follows that both belong to the same family of similar matrices. However, any matrix A and A^T can be shown to be similar to the Jordan normal form upper-triangular matrix and its transpose.

17. Let \bar{x} be a right eigenvector (column vector) of square matrix A with eigenvalue λ_r . Let \bar{y} be a left eigenvector (row vector) of A with eigenvalue $\lambda_l \neq \lambda_r$. Show that \bar{x} and \bar{y}^T are orthogonal.

$$\underbrace{\bar{y} [A \bar{x}]}_{\lambda_r \bar{y} \bar{x}} = \underbrace{[\bar{y} A]}_{\lambda_l \bar{y}} \bar{x}$$

$$\lambda_r \bar{y} \bar{x} = \lambda_l \bar{y} \bar{x}$$

Since the two eigenvalues are different, the above condition is possible only when $\bar{y} \bar{x} = 0$.

18. True or False? (a) A matrix with all zero eigenvalues must be the zero matrix. (b) A symmetric matrix with all zero eigenvalues must be the zero matrix.

(a) False. Any nilpotent strictly triangular matrix has all zero eigenvalues, but it is not the zero matrix. It is also not diagonalizable. (b) True. The key point is that symmetric matrices are also diagonalizable. Therefore, it has to be in the form $P\Delta P^T$, where Δ is the zero matrix. Therefore, the matrix is also a zero matrix.

19. Show that if λ is a non-zero eigenvalue of AB , then it must also be a non-zero eigenvalue of BA . Why does this argument not work for zero eigenvalues? Furthermore, show that if either A or B is invertible, then AB and BA are similar.

Let \bar{x} and λ be an eigenvector-eigenvalue pair of AB . Then, we have $AB\bar{x} = \lambda\bar{x}$. Left multiplying both sides with B we obtain $BAB\bar{x} = \lambda B\bar{x}$. One can regroup as $BA[B\bar{x}] = \lambda B\bar{x}$. Therefore, $B\bar{x}$ is an eigenvector of BA with eigenvalue λ . In the case of zero eigenvalues we have $AB\bar{x} = \bar{0}$. Therefore, we have $BA[B\bar{x}] = \bar{0}$. Unfortunately, we have no way of knowing whether $B\bar{x}$ is the zero vector. Eigenvectors are defined as nonzero vectors.

When B is invertible, the two matrices are similar because we have $BA = B(AB)B^{-1}$. This is precisely the similarity relationship. A similar argument can be made when A is invertible.

20. Is the quadratic function $f(x_1, x_2, x_3) = 2x_1^2 + 3x_2^2 + 2x_3^2 - 3x_1x_2 - x_2x_3 - 2x_1x_3$ convex? How about the function $g(x_1, x_2, x_3) = 2x_1^2 - 3x_2^2 + 2x_3^2 - 3x_1x_2 - x_2x_3 - 2x_1x_3$? In each case, find the minimum of the objective function, subject to the constraint that the norm of $[x_1, x_2, x_3]^T$ is 1.

The first quadratic function is convex because its 3×3 symmetric matrix can be shown to all have positive eigenvalues, which makes its positive semi-definite. The second quadratic function is not convex because its matrix has a negative diagonal entry. In each case, the smallest normalized eigenvector of the corresponding matrix provides the solution $[x_1, x_2, x_3]$.

21. Consider the function $f(x_1, x_2) = x_1^2 + 3x_1x_2 + 6x_2^2$. Propose a linear transformation of the variables so that the function is separable in terms of the new variables. Use the separable form of the objective function to find an optimal solution for the minimization problem.

One can use either eigenvectors or the generalized Gram-Schmidt method in order to create conjugate directions. In this particular case, the expression can even be represented as $(x_1 + 3x_2/2)^2 + 15x_2^2/4$. Therefore, the variable transformations are $y_1 = x_1 + 3x_2/2$ and $y_2 = 0$. As a result, one obtains $y_1 = y_2 = 0$, which in turn yields $x_1 = x_2 = 0$.

22. Show that the difference between two similar, symmetric matrices must be indefinite, unless both matrices are the same.

The difference between two symmetric matrices is symmetric, which makes the matrices diagonalizable. The trace of the matrix is zero, which means that the sum of eigenvalues of a diagonalizable matrix is 0. Therefore, some eigenvalues will have conflicting signs, unless the matrix is the zero matrix.

23. Show that an n th root of a $d \times d$ diagonalizable matrix can always be found, as long as we allow for complex eigenvalues. Provide a geometric interpretation of the resulting matrix in terms of its relationship to the original matrix in the case where the root is a real-valued matrix.

The n th root of $P\Delta P^{-1}$ is $P\Delta^{1/n}P^{-1}$. Note that the diagonal matrix might contain complex eigenvalues. The scaling factors in the different directions are now the n th roots along these directions.

24. Generate the equation of an ellipsoid centered at $[1, -1, 1]^T$, and whose axes directions are the orthogonal vectors $[1, 1, 1]^T$, $[1, -2, 1]^T$, and $[1, 0, -1]^T$. The ellipsoid is stretched in these directions in the ratio $1 : 2 : 3$. The answer to this question is not unique, and it depends on the size of your ellipsoid. Use the matrix form of ellipsoids discussed in the chapter.

Create a matrix P containing the *normalized* eigenvectors (axes directions in problem statement) in its columns. Let \bar{c} be the center of the ellipsoid. Let Δ be a diagonal matrix which contains the *inverse squares* of the scale factors in its diagonal entries. These factors are 1, $1/2^2$ and $1/3^2$. Then, the matrix A is $A = P\Delta P^T$. Then, the equation of the ellipsoid is $[\bar{x} - \bar{c}]^T A [\bar{x} - \bar{c}] = s^2$.

25. If A and B are symmetric matrices whose eigenvalues lie in $[\lambda_1, \lambda_2]$ and $[\gamma_1, \gamma_2]$, respectively, show that the eigenvalues of $A - B$ lie in $[\lambda_1 - \gamma_2, \lambda_2 - \gamma_1]$.

Let \bar{x} be a unit vector and $C = P|\Delta|P^T$. Then, we can show that $\bar{x}^T C \bar{x} = (P^T \bar{x})^T \Delta (P^T \bar{x}) = \bar{y}^T \Delta \bar{y}$, where \bar{y} is the new unit vector in the orthogonal basis system defined by P . The maximum value of $\bar{y}^T \Delta \bar{y}$ is the maximum entry of Δ and the minimum value is the minimum entry of Δ . Now recognize that $\bar{x}^T (A - B) \bar{x}$ can be written as $\bar{x}^T A \bar{x} - \bar{x}^T B \bar{x}$. For each of the individual terms use its minimum or maximum value to obtain the desired result.

26. Consider a nonzero, square matrix A satisfying $A^k = 0$ for some k . Show that all eigenvalues are 0 and such a matrix is defective.

Let λ be an eigenvalue. Then, for any eigenvector \bar{x} we have $A^k \bar{x} = \lambda^k \bar{x}$. This is possible only when $\lambda = 0$. If the matrix were to be diagonalizable, it would have a diagonal matrix containing 0s. This is possible only when A is a zero matrix, which is a contradiction.

27. Show that A is diagonalizable in each case if (i) it satisfies $A^2 = A$, and (ii) it satisfies $A^2 = I$.

(i) In this case, the eigenvalues are either 1 or 0. Consider the case where the eigenvalue with repeated multiplicity is 1. If the matrix is not diagonalizable because of this eigenvalue, we can find a Jordan chain containing \bar{x} and \bar{y} so that $A\bar{x} = \bar{x} + \bar{y}$, and $A\bar{y} = \bar{y}$. Multiplying the first equation with A , we get $A^2\bar{x} = \bar{x} + 2\bar{y}$. However, since $A^2 = A$, it follows that \bar{y} must be the zero vector. One can make a similar argument about the eigenvalue of 0.

(ii) This part is similar to (i), except that the eigenvalues are 1 or -1 .

28. **Elementary Row Addition Matrix Is Defective:** Show that the $d \times d$ elementary row addition matrix with 1s on the diagonal and a single nonzero off-diagonal entry is not diagonalizable.

Let a_{ij} be the non-zero off diagonal entry. The characteristic polynomial is $(1 - \lambda)^d$, which yields a single eigenvalue of 1 with algebraic multiplicity d . However, the geometric multiplicity of this eigenvalue is $(d - 1)$, since $a_{ij}x_j = 0$, and we cannot choose any vector with $x_j = 1$ in this eigenspace. Therefore, the matrix is defective.

29. Show that any $n \times n$ matrix P satisfying $P^2 = P$ and $P = P^T$ can be expressed in the form QQ^T for some $n \times d$ matrix Q with orthogonal columns (and is hence a projection matrix).

Since the matrix is symmetric, it can be diagonalized as $O\Sigma O^T$ for some orthogonal matrix O . Also, according to the solution to problem 27, the eigenvalues of this matrix are either 1 or 0. Therefore, one can drop the eigenvectors with zero eigenvalues and the diagonal entries of zero to write the matrix equivalently as $Q\Sigma Q^T$, where Q is a rectangular matrix, and Σ is a smaller diagonal matrix containing only 1s (which is a smaller identity matrix). Therefore, the matrix can be written as QQ^T .

30. **Diagonalizability and Nilpotency:** Show that every square matrix can be expressed as the sum of a diagonalizable matrix and a nilpotent matrix (including zero matrices for either part).

The Jordan normal form provides the mechanism to do this. Any matrix A can be represented in the following form $A = PJP^{-1}$. The matrix $J = \Delta + T$, where Δ is diagonal and T is strictly triangular (nilpotent). Furthermore, we can express the Jordan normal form as $P\Delta P^{-1} + PTP^{-1}$. The first part is diagonalizable and the second part is nilpotent.

31. Suppose you are given the Cholesky factorization LL^T of a positive definite matrix A . Show how to compute the inverse of A using multiple applications of back substitution.

Suppose that X is the inverse of the $d \times d$ matrix A . Therefore, we have $LL^T X = I$. Let $Y = L^T X$. Therefore, we have $LY = I$. Therefore, we have the following:

$$L[\bar{y}_1 \dots \bar{y}_d] = [\bar{e}_1 \dots \bar{e}_d]$$

Therefore, we have $L\bar{y}_i = \bar{e}_i$. We can solve for \bar{y}_i using backsubstitution. One we have solved for Y , we have the second system of equations:

$$L^T X = L^T[\bar{x}_1 \dots \bar{x}_d] = [\bar{y}_1 \dots \bar{y}_d]$$

Therefore we have $L^T \bar{x}_i = \bar{y}_i$. Since \bar{x}_i is known now for each i , we can now solve for each \bar{x}_i . Therefore, the inverse can be constructed as $X = [\bar{x}_1 \dots \bar{x}_d]$.

32. **Rotation with arbitrary axis:** Suppose that the vector $[1, 2, -1]^T$ is the axis of a counter-clockwise rotation of θ degrees, just as $[1, 0, 0]^T$ is the axis of the counter-clockwise θ -rotation of the following Givens matrix:

$$R_{[1,0,0]} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix}$$

Create a new orthogonal basis system of \mathcal{R}^3 that includes $[1, 2, -1]^T$. Now use the concept of similarity $R_{[1,2,-1]} = PR_{[1,0,0]}P^T$ to create a 60° rotation matrix M about the axis $[1, 2, -1]^T$. The main point is in knowing how to infer P from the aforementioned orthogonal basis system. Be careful of the handedness of the new axis system. Now show how to recover the axis and angle of rotation from M using complex-valued diagonalization.

One can pick any pair of orthogonal vectors of $[1, 2, -1]$, which are $[1, 0, 1]$, and $[1, -1, -1]$, respectively. These vectors need to be normalized and put in the columns

of P . Furthermore, to make sure that the matrix is a pure rotation without reflection, the determinant of the matrix needs to be 1 rather than -1 . Keeping the columns in the aforementioned order ensures that their determinant is $+1$ rather than -1 . Therefore, we obtain the following matrix:

$$P = \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ 2/\sqrt{6} & 0 & -1/\sqrt{3} \\ -1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \end{bmatrix}$$

The Givens rotation matrix for a 60° counter-clockwise rotation of column vectors is as follows:

$$R_{60} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(60) & -\sin(60) \\ 0 & \sin(60) & \cos(60) \end{bmatrix}$$

Note that the ordering of the columns of the Givens rotation and the matrix P need to be such that the axis column corresponds to the column $[1, 0, 0]^T$ in the rotation matrix. This ensures that the coordinates lying on the axis of rotation do not change. Then, the overall rotation matrix is as follows:

$$M = PR_{60}P^T = \begin{bmatrix} 0.58333333 & 0.52022006 & 0.62377345 \\ -0.18688672 & 0.83333333 & -0.52022006 \\ -0.79044011 & 0.18688672 & 0.58333333 \end{bmatrix}$$

Just to confirm that this is indeed a 60° rotation matrix, we can pick any vector orthogonal to the axis of rotation and check if it is rotated by 60° . For example, let us test the unit vector $\bar{x} = [-3, 1, -1]^T/\sqrt{11}$. It is easy to verify that the transformed version of this unit vector with matrix M is the following:

$$M\bar{x} = \frac{1}{\sqrt{11}} \begin{bmatrix} -1.85355339 \\ 1.91421356 \\ 1.97487373 \end{bmatrix}$$

The dot product between \bar{x} and $M\bar{x}$ can be shown to be 0.5, which is the cosine of 60° .

Next, we show the reverse process of extracting the rotation angles and the axis of rotation from matrix M . We first extract the eigenvectors and eigenvalues of M .

On performing the diagonalization of M (use any package such as numpy), we obtain that the eigenvalues are 1, $\cos(60) + i\sin(60)$, and $\cos(60) - i\sin(60)$. The rotation is clearly 60° , but we do not yet know whether this is clockwise or counter-clockwise. The invariant eigenvector is returned as $[-1, -2, 1]^T$. This is the axis of rotation, although we note that it has been multiplied by -1 during recovery. This type of ambiguity is natural during recovery because more than one answer is correct. Because of this multiplication of the axis by -1 , the rotation occurs 60° *clockwise* with respect to this axis. This fact can be verified by creating the right-handed basis $[-1, -2, 1]$, $[1, 0, 1]$, and $[-1, 1, 1]$, and testing the final resting position $M\bar{z}$ of the second basis vector $\bar{z} = [1, 0, 1]^T$ on performing the transformation $M\bar{z}$. A 60° **clockwise** rotation will result in a **negative** dot product of $M\bar{z} = [1.21, -0.71, -0.21]^T$ with the third basis direction $[-1, 1, 1]^T$. It is evident that this is indeed the case here.

- 33.** Suppose that you are given the Jordan normal form of a matrix. Show how you can use this form to quickly identify the rank of the matrix and all four fundamental subspaces of the matrix.

Let the Jordan normal form be PJP^{-1} . Then, the columns of P corresponding to the non-zero eigenvectors (ordinary or generalized) together with the generalized zero eigenvectors form the column space. A similar statement can be made about the row space using rows of P^{-1} . The orthogonal complement of the column space is the left null space (zero ordinary eigenvectors in P), and the orthogonal complement of the row space is the right null space (zero left eigenvectors of P^{-1}).

- 34.** Consider the following quadratic form:

$$f(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + x_3^2 + ax_1x_2 + x_2x_3$$

Under what conditions on a is the function $f(x_1, x_2, x_3)$ convex?

Let H be the Hessian. Then, the value of $H - \lambda I$ is computed as follows:

$$H = \begin{bmatrix} 2 - \lambda & a & 0 \\ a & 4 - \lambda & 1 \\ 0 & 1 & 2 - \lambda \end{bmatrix}$$

Let us compute the characteristic polynomial of the Hessian.

$$\begin{aligned} f(\lambda) &= (2 - \lambda)[(4 - \lambda)(2 - \lambda) - 1] - a^2(2 - \lambda) \\ &= (2 - \lambda)[(4 - \lambda)(2 - \lambda) - 1] - a^2(2 - \lambda) \\ &= (2 - \lambda)(7 - 6\lambda + \lambda^2) - a^2(2 - \lambda) \\ &= (14 - 2a^2) - (19 - a^2)\lambda + 2\lambda^2 - \lambda^3 \end{aligned}$$

All roots of this equation need to be nonnegative. The constant coefficient is the product of the roots, which needs to be non-negative. Therefore, we have $14 - 2a^2 \geq 0$, which implies that $|a| \leq \sqrt{7}$. Second, the pairwise product of the roots must be positive. Therefore, we have $19 - a^2 \geq 0$. Therefore, we have $|a| \leq \sqrt{19}$. This inequality is subsumed by the first one. Therefore, we need $|a| \leq \sqrt{7}$.

- 35.** Consider an $n \times n$ non-singular matrix $A = BB^T$, which is the left Gram Matrix of B . Propose an algorithm that takes B as input and generates 100 different matrices, $B_1 \dots B_{100}$, such that A is the left Gram matrix of each B_i . How many such matrices exist? Is it possible to obtain a B_i that is also symmetric like A ? Is any B_i triangular?

Any matrix $B_i = BP$ for orthogonal matrix P would be such that A is a left Gram matrix of B_i . One can therefore generate 100 different orthogonal matrices by using Gram-Schmidt orthogonalization on random vectors. One of the possible matrices is the square-root matrix, which is symmetric. Cholesky factorization yields a triangular B .

- 36.** Let P be an $n \times n$ nonnegative stochastic transition matrix of probabilities, so that the probabilities in each row sum to 1. Find a right eigenvector with eigenvalue 1 by inspection. Prove that no eigenvalue can be larger than 1.

The vector of 1s, is an eigenvector with eigenvalue 1. Because of the fact that each row of P sums to 1, the result of $P\bar{x}$ creates a vector \bar{y} , so that each y_i is a weighted

average of the various x_i . Therefore, none of the y_i can be strictly greater than the largest x_i . However, if we have an eigenvalue greater than 1, this condition will be violated.

- 37.** Suppose that $A = V\Delta V^{-1}$ is a diagonalizable matrix. Show that the matrix $\lim_{n \rightarrow \infty} (I + A/n)^n$ exists with finite entries.

This limit is equivalent to the following:

$$B = V \lim_{n \rightarrow \infty} (I + \Delta/n)^n V^{-1}$$

The limit would then be applied to each diagonal entry, which does exist. The resulting diagonal entries are $\exp(\lambda_1) \dots \exp(\lambda_d)$.

- 38.** Eigenvalues are scaling factors along specific directions. Construct an example of a 2×2 diagonalizable matrix A and 2-dimensional vector \bar{x} , so that each eigenvalue of A is less than 1 in absolute magnitude and the length of $A\bar{x}$ is larger than that of \bar{x} . Prove that any such matrix A cannot be symmetric. Provide an intuitive geometric explanation of both phenomena.

The matrix A and vector \bar{x} is defined as follows:

$$A = \begin{bmatrix} 0.9 & 0 \\ 0.9 & 0 \end{bmatrix}, \quad \bar{x} = \begin{bmatrix} 0.9 \\ 0 \end{bmatrix}, \quad A\bar{x} = \begin{bmatrix} 0.9 \\ 0.9 \end{bmatrix}$$

The eigenvalues of A are 0.9 and 0.

Now consider the symmetric matrix $P\Delta P^T$. We have the following:

$$\begin{aligned} \|A\bar{x}\| &= \|P(\Delta P^T \bar{x})\| = \|\Delta(P^T \bar{x})\| \quad [\text{Orthogonality of } P] \\ &< \|P^T \bar{x}\| \quad [\text{Diagonal } \Delta \text{ with each value less than 1}] \\ &= \|\bar{x}\| \quad [\text{Orthogonality of } P^T] \end{aligned}$$

The reason that eigenvalue can be less than 1 in asymmetric matrix but particular vectors are scaled at factors greater than 1 is that the directions of scaling are non-orthogonal. Therefore, the coordinates of a point can have larger norm than in the standard basis. Therefore, the overall scaling factor in an arbitrary direction is not guaranteed to be less than 1 even after reducing the lengths of individual components. On the other hand, in the case of symmetric matrix the directions of scaling are orthogonal; the coordinates of a point have the same norm as the point. Therefore, if \bar{x} is represented in the basis of P , both components reduce. Therefore, the vector length will always decrease.

- 39. Mahalanobis distance:** Let $C = D^T D/n$ be the covariance matrix of an $n \times d$ mean-centered data set. The squared Mahalanobis distance of the i th row \bar{X}_i of D to the mean of the data set (which is the origin in this case) is given by the following:

$$\delta_i^2 = \bar{X}_i C^{-1} \bar{X}_i^T$$

Let $C = P\Delta P^T$ be the diagonalization of C , and each row of \bar{X}_i be transformed to $\bar{Z}_i = \bar{X}_i P$. Normalize each attribute of the transformed data by dividing with the standard derivation (of the transformed data) to make its variance 1 along each

dimension and to create the new rows $\bar{Z}'_1 \dots \bar{Z}'_n$. Show that the Mahalanobis distance δ_i is equal to $\|\bar{Z}'_i\|$.

The value δ_i^2 can be written as follows:

$$\delta_i^2 = \bar{X}_i C^{-1} \bar{X}_i^T = \bar{X}_i (P \Delta P^T)^{-1} \bar{X}_i^T = (\bar{X}_i P) \Delta^{-1} (\bar{X}_i P)^T = \bar{Z}_i \Delta^{-1} \bar{Z}_i^T = \|\bar{Z}_i \Delta^{-1/2}\|^2$$

The key point is to understand that $\bar{Z}_i \Delta^{-1/2}$ contains the data after normalization with standard deviation. This is because Δ contains the variances of the transformed data. Therefore, $\bar{Z}_i \Delta^{-1/2}$ is the same as \bar{Z}'_i .

- 40. Non-orthogonal diagonalization of symmetric matrix:** Consider the following diagonalization of a symmetric matrix:

$$\begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix}$$

Find an alternative diagonalization $V \Delta V^{-1}$ in which the columns of V are not orthogonal.

The key point here is that one can choose any basis of tied eigenvectors to construct V and then calculate V^{-1} after the fact. For example, one can modify the above diagonalization by adding columns 1 and 2 of the orthogonal matrix to create V with a new column 1. Then, the diagonalization is as follows:

$$\begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1 & 1 & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/\sqrt{2} & 1 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{bmatrix}$$

Note that this type of non-orthogonal diagonalization is not possible if there are no tied eigenvalues.

- 41. Power method with Gram matrix:** Suppose that you have a 100000×100 sparse matrix D , and you want to compute the dominant eigenvector of the left Gram matrix DD^T . Unfortunately, DD^T is of size 100000×100000 , and it might not even be sparse. This can cause computational problems. Show how you can implement the power method using only sparse matrix-vector multiplications.

While implementing the power iterations, perform the updates as follows:

$$\bar{x} \leftarrow \frac{D(D^T \bar{x})}{\|D(D^T \bar{x})\|}$$

Note the nesting of the brackets, which makes sure that DD^T is never materialized.

- 42. Multiple choice:** Suppose $\bar{x}_i^T A \bar{x}_i > 0$ for d vectors $\bar{x}_1 \dots \bar{x}_d$ and $d \times d$ symmetric matrix A . Then, A is positive definite if the different \bar{x}_i 's are (i) linearly independent, (ii) orthogonal, (iii) A -orthogonal, (iv) any of the above, or (v) none of the above? Justify your answer.

The correct answer is choice (iii). This is because if the vectors are A -orthogonal and each $\bar{x}_i^T A \bar{x}_i > 0$, then the vectors $\bar{x}_1 \dots \bar{x}_d$ are linearly independent. This follows from the solution of Exercise 43 of the previous chapter. One can express any vector $\bar{x} = \sum_{i=1}^d \alpha_i \bar{x}_i$. Plugging this value of \bar{x} in $\bar{x}^T A \bar{x}$, we obtain $\sum_{i=1}^d \alpha_i^2 \bar{x}_i^T A \bar{x}_i$, which is greater than 0.

- 43.** Convert the diagonalization in the statement of Exercise 40 into Gram matrix form $A = B^T B$ and then compute the Cholesky factorization $A = LL^T = R^T R$ using the QR decomposition $B = QR$.

For a diagonalization $A = Q\Delta Q^T$, we can convert it into a symmetric factorization $A = B^T B$ using $B = (Q\sqrt{\Delta})^T$. In this case, the matrix B turns out to be the following:

$$B = \begin{bmatrix} \sqrt{2} & 0 & \sqrt{2} \\ 0 & 2 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

Next, we perform the QR decomposition of B as follows:

$$B = \begin{bmatrix} \sqrt{2} & 0 & \sqrt{2} \\ 0 & 2 & 0 \\ 1 & 0 & -1 \end{bmatrix} = QR = \begin{bmatrix} \sqrt{2/3} & 0 & -\sqrt{1/3} \\ 0 & 1 & 0 \\ \sqrt{1/3} & 0 & \sqrt{2/3} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & \sqrt{1/3} \\ 0 & 2 & 0 \\ 0 & 0 & -\sqrt{8/3} \end{bmatrix}$$

The matrix L is simply the transpose of R , which is the following:

$$L = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 2 & 0 \\ \sqrt{1/3} & 0 & -\sqrt{8/3} \end{bmatrix}$$

The Cholesky factorization is simply LL^T . Note that the Cholesky factorization can be derived from any symmetric decomposition of a matrix by simply using the QR method on the factor matrix. This point is also made in one of the practice problems in the section on Cholesky factorization.

Chapter 4

Optimization Basics: A Machine Learning View

1. Find the saddle points, minima, and the maxima of the following functions:

(a) $F(x) = x^2 - 2x + 2$

(b) $F(x, y) = x^2 - 2x - y^2$

For (a) $x = 1$ is a minimum, as it satisfies both first-order and second-order condition.
(b) $x = 1$ and $y = 0$ is a critical point with an indefinite Hessian.

2. Suppose that \bar{y} is a d -dimensional vector with very small norm $\epsilon = \|\bar{y}\|_2$. Consider a continuous and differentiable objective function $J(\bar{w})$ with zero gradient and Hessian H at $\bar{w} = \bar{w}_0$. Show that $\bar{y}^T H \bar{y}$ is approximately equal to twice the change in $J(\bar{w})$ by perturbing $\bar{w} = \bar{w}_0$ by ϵ in direction $\bar{y}/\|\bar{y}\|$.

This result can be shown by using the second-order Taylor series expansion about \bar{w} .

3. Suppose that an optimization function $J(\bar{w})$ has a gradient of 0 at $\bar{w} = \bar{w}_0$. Furthermore, the Hessian of $J(\bar{w})$ at $w = \bar{w}_0$ has both positive and negative eigenvalues. Show how you would use the Hessian to (i) find a vector direction along which infinitesimal movements in either direction from \bar{w}_0 decrease $J(\bar{w})$; (ii) find a vector direction along which infinitesimal movements in either direction from \bar{w}_0 increase $J(\bar{w})$. Is \bar{w}_0 is maximum, minimum, or saddle-point?

The point is a saddle point. Positive eigenvectors correspond to directions along which the point is a local minimum, whereas negative eigenvectors correspond to directions along which the point is a local maximum.

4. We know that the maximum of two convex functions is a convex functions. Is the minimum of two convex functions convex? Is the intersection of two convex sets convex? If the union of two convex sets convex? Justify your answer in each case.

Minimum is not convex. For example, the minimum of $|x|$ and $|x+2|$ is not convex with two global minima and $x = 0, -2$. Intersection is convex and can be shown by using the convexity condition. Union of two convex functions is not convex. For example, the union of all the points lying on two straight lines in 2 dimensions is not a convex set.

5. Either prove each statement or give a counterexample: (i) If $f(x)$ and $g(x)$ are convex, then $F(x, y) = f(x) + g(y)$ is convex. (ii) If $f(x)$ and $g(x)$ are convex, then $F(x, y) = f(x) \cdot g(y)$ is convex.

The first statement is true. It is very similar to the proof that the sum of convex functions is convex (although this is a bivariate function created out of two univariate functions). The second statement is not true. For example, $f(x) = x$ and $g(x) = x$ are convex, but $F(x, y) = xy$ is not convex.

6. **Hinge-loss without margin:** Suppose that we modified the hinge-loss by removing the constant value within the maximization function as follows:

$$J = \sum_{i=1}^n \max\{0, (-y_i[\bar{W} \cdot \bar{X}_i])\} + \frac{\lambda}{2} \|\bar{W}\|^2$$

This loss function is referred to as the perceptron criterion. Derive the stochastic gradient descent updates for this loss function.

The stochastic gradient descent updates are identical to the perceptron, except that we perform updates for any point lying on the wrong side of the decision boundary (and not worrying about the correctly classified points that are too close to the decision boundary).

7. Compare the perceptron criterion of the previous exercise to the hinge-loss in terms of its sensitivity to the magnitude of \bar{W} . State one non-informative weight vector \bar{W} , which will always be an optimal solution to the optimization problem of the previous exercise. Use this observation to explain why a perceptron (without suitable modifications) can sometimes provide much poorer solutions with an SVM when the points of the two classes cannot be separated by a linear hyperplane.

A zero weight vector provides a loss value of 9. Therefore, when the points are inseparable, the perceptron algorithm will focus too much on reducing the magnitude of the weight vector.

8. Consider an unconstrained quadratic program of the form $\bar{w}^T A \bar{w} + \bar{b}^T \bar{w} + c$, where \bar{w} is a d -dimensional vector of optimization variables, and the $d \times d$ matrix A is positive semi-definite. The constant vector \bar{b} is d -dimensional. Show that a global minimum exists for this quadratic program if and only if \bar{b} lies in the column space of A .

The gradient of this objective function is $2A\bar{w} + \bar{b}$. In other words, the system $A\bar{w} = -\bar{b}/2$ needs to have a solution. This is possible if and only if \bar{b} lies in the column space of A .

One can show an unbounded solution more explicitly by using a variable transformation that creates a separable objective function (as discussed in section 3.3.3). With this type of variable transformation, the linear part ends up having variables that are not present in the quadratic part, and therefore those unmatched variables can be set to extremely large positive or negative values to create an unbounded objective function value. Let $A = Q\Delta Q^T$ be the diagonalization of A . We only need to consider the case where A is not of full rank, and therefore some of the eigenvalues of A will be 0. In such a case, one can perform the transformation as $A = P\Delta P^T$, and define the new set of variables $\bar{v} = P^T \bar{w}$. Then, the objective function can be written as follows:

$$J = \bar{v}^T \Delta \bar{v} + \bar{b}^T P \bar{v} + c$$

Note that the column space of A is defined by the nonzero eigenvectors of A . Therefore $\bar{d} = P^T \bar{b}$ will be a vector in which there will be at least one nonzero entry d_j at $j = j_1$ for which $\Delta_{jj} = 0$. The linear part of the objective function is of the form $\sum_r d_r v_r$, and the quadratic part is of the form $\delta_{rr} \|\bar{v}_r\|^2$. Depending on whether d_{j_1} is positive or negative we can set v_{j_1} to unboundedly negative or positive values.

9. *The text of the book discusses a stochastic gradient descent update of the Weston-Watkins SVM, but not a mini-batch update. Consider a setting in which the mini-batch S contains training pairs of the form (\bar{X}, c) , where each $c \in \{1, \dots, k\}$ is the categorical class label. Show that the stochastic gradient-descent step for each separator \bar{W}_r at learning rate α :*

$$\bar{W}_r \leftarrow \bar{W}_r(1 - \alpha\lambda) + \alpha \sum_{(\bar{X}, c) \in S, r=c} \bar{X}^T [\sum_{j \neq r} \delta(j, \bar{X})] - \alpha \sum_{(\bar{X}, c) \in S, r \neq c} \bar{X}^T [\delta(r, \bar{X})] \quad (4.1)$$

Here, \bar{W}_r is defined in the same way as the text of the chapter.

Let $J(S)$ be the objective function defined only over the mini-batch S as follows:

$$J(S) = \sum_{(\bar{X}, c) \in S} \sum_{r: r \neq c(i)} \max(\bar{W}_r^T \cdot \bar{X} - \bar{W}_c^T \cdot \bar{X} + 1, 0) + \frac{\lambda}{2} \sum_{r=1}^k \|\bar{W}_r\|^2$$

The loss for each training instance can be decomposed into its loss from the different separators. On working out the details of the gradient with respect to the parameters in r th separator \bar{W}_r , we obtain the following:

$$\frac{\partial J(S)}{\partial \bar{W}_r} = \lambda \bar{W}_r - \sum_{(\bar{X}, c) \in S, r=c} \bar{X}^T [\sum_{j \neq r} \delta(j, \bar{X})] + \sum_{(\bar{X}, c) \in S, r \neq c} \bar{X}^T [\delta(r, \bar{X})] \quad (4.2)$$

This results in the following stochastic gradient-descent step for each separator \bar{W}_r at learning rate α :

$$\bar{W}_r \leftarrow \bar{W}_r(1 - \alpha\lambda) + \alpha \sum_{(\bar{X}, c) \in S, r=c} \bar{X}^T [\sum_{j \neq r} \delta(j, \bar{X})] - \alpha \sum_{(\bar{X}, c) \in S, r \neq c} \bar{X}^T [\delta(r, \bar{X})] \quad (4.3)$$

10. *Consider the following function $f(x, y) = x^2 + 2y^2 + axy$. For what values of a (if any) is the function $f(x, y)$ concave, convex, and indefinite?*

The Hessian of this function is as follows:

$$H = \begin{bmatrix} 2 & a \\ a & 4 \end{bmatrix}$$

The characteristic polynomial of this matrix is $(2 - \lambda)(4 - \lambda) - a^2$. This polynomial simplifies to $\lambda^2 - 6\lambda + 8 - a^2$. Both roots of this equation can never be non-positive, since the sum of the roots is 6. So the Hessian is never negative semi-definite, and the function is not concave. However, the product of the roots is $8 - a^2$. Therefore, both roots are positive if and only if $8 - a^2 > 0$. In other words, the function is convex for $a \in [-\sqrt{8}, \sqrt{8}]$, and it is indefinite otherwise.

11. Consider the bivariate function $f(x, y) = x^3/6 + x^2/2 + y^2/2 + xy$. Define a domain of values of the function, at which it is convex.

The Hessian of this function is as follows:

$$H = \begin{bmatrix} 1+x & 1 \\ 1 & 1 \end{bmatrix}$$

The characteristic polynomial is $\lambda^2 - (2+x)\lambda$. Therefore the roots are $\lambda = 0$ and $\lambda = 2+x$. In order for the function to be convex all eigenvalues have to be nonnegative. Therefore, we have $x \geq -2$. Therefore, restricting the domain of this function to $x \geq -2$ yields a convex function.

12. Consider the L_1 -loss function for binary classification, where for feature-class pair (\bar{X}_i, y_i) and d -dimensional parameter vector \bar{W} , the point-specific loss for the i th instance is defined as follows:

$$L_i = \|y_i - \bar{W} \cdot \bar{X}_i^T\|_1$$

Here, we have $y_i \in \{-1, +1\}$, and \bar{X}_i is a d -dimensional row vector of features. The norm used above is the L_1 -norm instead of the L_2 -norm of least-squares classification. Discuss why the loss function can be written as follows for $y_i \in \{-1, +1\}$:

$$L_i = \|1 - y_i \bar{W} \cdot \bar{X}_i^T\|_1$$

Show that the stochastic gradient descent update is as follows:

$$\bar{W} \leftarrow \bar{W}(1 - \alpha\lambda) + \alpha y_i \bar{X}_i^T \text{sign}(1 - y_i \bar{W} \cdot \bar{X}_i^T)$$

Here, λ is the regularization parameter, and α is the learning rate. Compare this update with the hinge-loss update for SVMs.

Since, $y_i^2 = 1$, the loss function can be written as follows:

$$L_i = \|y_i - y_i^2 \bar{W} \cdot \bar{X}_i^T\|_1 = \underbrace{|y_i|}_1 \|1 - y_i \bar{W} \cdot \bar{X}_i^T\|_1 = \|1 - y_i \bar{W} \cdot \bar{X}_i^T\|_1$$

The differentiation of the modulus is the sign operator. Therefore, we have the following:

$$\frac{\partial L_i}{\partial \bar{W}} = -y_i \bar{X}_i^T \text{sign}(1 - y_i \bar{W} \cdot \bar{X}_i^T)$$

This results in the following regularized stochastic gradient descent update:

$$\bar{W} \leftarrow \bar{W}(1 - \alpha\lambda) - \alpha \frac{\partial L_i}{\partial \bar{W}}$$

The result follows. The hinge-loss can be viewed as a “repair” of this function, where overperformance is not penalized. Therefore, when the value of $1 - y_i \bar{W} \cdot \bar{X}_i^T < 0$, the update is not performed for hinge-loss SVM. However, in this case, the update will still be performed in the case of L_1 -loss. This is the only difference from the hinge-loss SVM. This is similar to the relationship between least-squares classification and the L_2 -SVM.

13. Let \bar{x} be an n_1 -dimensional vector, and W be an $n_2 \times n_1$ -dimensional matrix. Show how to use the vector-to-vector chain rule to compute the vector derivative of $(W\bar{x}) \odot (W\bar{x})$ with respect to \bar{x} . Is the resulting vector derivative a scalar, vector, or matrix? Now repeat this exercise for $F((W\bar{x}) \odot (W\bar{x}))$, where $F(\cdot)$ is a function summing the elements of its argument into a scalar.

Let $\bar{o} = \bar{h} \odot \bar{h}$ and $\bar{h} = W\bar{x}$. Then, the vector derivative is as follows:

$$\frac{\partial \bar{o}}{\partial \bar{x}} = \frac{\partial \bar{h}}{\partial \bar{x}} \frac{\partial \bar{o}}{\partial \bar{h}} = W^T (2\Delta)$$

Here, Δ is a diagonal matrix in which the k th diagonal entry contains the k th entry of \bar{h} .

In the second case, the final scalar value is $J = F(\bar{o})$. In such a case, the chain rule tells us the following:

$$\frac{\partial J}{\partial \bar{x}} = \frac{\partial \bar{h}}{\partial \bar{x}} \frac{\partial \bar{o}}{\partial \bar{h}} \frac{\partial J}{\partial \bar{o}} = W^T (2\Delta) \bar{1} = 2W^T \bar{h} = 2W^T W\bar{x}$$

14. Let \bar{x} be an n_1 -dimensional vector, and W be an $n_2 \times n_1$ -dimensional matrix. Show how to use the vector-to-vector chain rule to compute the vector derivative of $W(\bar{x} \odot \bar{x} \odot \bar{x})$ with respect to \bar{x} . Is the resulting vector derivative a scalar, vector, or matrix? Now repeat this exercise for $G(W(\bar{x} \odot \bar{x} \odot \bar{x}) - \bar{y})$, where \bar{y} is a constant vector in n_2 -dimensions, and $G(\cdot)$ is a function summing the absolute value of the elements of its argument into a scalar. You may find it helpful to express $G(\cdot)$ as a composition of functions.

The derivative is a matrix. One can first create the vector-to vector function $F(\bar{x}) = \bar{x} \odot \bar{x} \odot \bar{x}$. Now define $H(\bar{x}) = WF(\bar{x})$. The vectored derivative of $F(\bar{x})$ is a diagonal matrix Δ of size $n_1 \times n_1$, in which the (i, i) th entry is $3x_i^2$. By using the vectored chain rule, the derivative of $H(\bar{x})$ with respect to \bar{x} is ΔW^T . Note the derivative of $G(\cdot)$ with respect to its argument is a column vector containing the sign of its argument. By further using the chain rule, one obtains the derivative of the function $G(\cdot)$ with respect to \bar{x} as $\Delta W^T \text{sign}[W(\bar{x} \odot \bar{x} \odot \bar{x}) - \bar{y}]$. Here, the sign function is applied in element-wise fashion to its vector argument.

15. Show that if scalar L can be expressed as $L = f(W\bar{x})$ for $m \times d$ matrix W and d -dimensional vector \bar{x} , then $\frac{\partial L}{\partial W}$ will always be a rank-1 matrix or a zero matrix irrespective of the choice of function $f(\cdot)$.

Let $\bar{h} = W\bar{x} = \sum_i x_i \bar{w}_i$, where \bar{w}_i is the i th column of W . Then, using the vector-to-vector chain rule, we obtain the following:

$$\frac{\partial L}{\partial \bar{w}_i} = \frac{\partial \bar{h}}{\partial \bar{w}_i} \frac{\partial L}{\partial \bar{h}} = x_i I_m \frac{\partial L}{\partial \bar{h}} = x_i \frac{\partial L}{\partial \bar{h}}$$

Note that $\frac{\partial \bar{h}}{\partial \bar{w}_i}$ is an $m \times m$ diagonal matrix in which every diagonal entry is x_i , and can therefore be expressed as $x_i I_m$, where I_m is an $m \times m$ identity matrix. The partial derivative of L with respect to W can be obtained by stacking up all these derivatives in the d columns. This leads to the outer-product form:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \bar{h}} \bar{x}^T$$

An outer-product matrix is either a rank-1 matrix or the zero matrix. The result follows.

- 16. Incremental linear regression with added points:** Suppose that you have a data matrix D and target vector \bar{y} in linear regression. You have done all the hard work to invert $(D^T D)$ and then compute the closed-form solution $\bar{W} = (D^T D)^{-1} D^T \bar{y}$. Now you are given an additional training point (\bar{X}, y) , and are asked to compute the updated parameter vector \bar{W} . Show how you can do this efficiently without having to invert a matrix from scratch. Use this result to provide an efficient strategy for incremental linear regression.

Whenever a row vector \bar{X} is appended to the bottom of matrix D , the corresponding matrix $D^T D$ has the same size, but the matrix $\bar{X} \bar{X}^T$ gets added to it. This is a rank-1 matrix. The efficient incremental inversion of such a matrix can be done using the matrix inversion lemma of Chapter 1.

- 17. Incremental linear regression with added features:** Suppose that you have a data set with a fixed number of points, but with an ever-increasing number of dimensions (as data scientists make an ever-increasing number of measurements and surveys). Provide an efficient strategy for incremental linear regression with regularization.

The solution to linear regression with regularization can be expressed equivalently as either $(D^T D + \lambda I_d)^{-1} D^T \bar{y}$ or is $D^T (D D^T + \lambda I_n)^{-1} \bar{y}$. The matrix $D D^T + \lambda I_n$ gets updated by $\bar{c} \bar{c}^T$ whenever a column vector is added to D . This is a rank-1 update that can be handled with the matrix inversion lemma.

- 18. Frobenius norm to matrix derivative:** Let A be an $n \times d$ constant matrix and V be a $d \times k$ matrix of parameters. Let \bar{v}_i be the i th row of V and \bar{V}_j be the j th column of V . Show the following:

- The matrix $\frac{\partial J}{\partial V}$ be written by stacking up the row vectors $\frac{\partial J}{\partial \bar{v}_i}$ into a matrix. How would you do this using columns of V ? These tricks enable the use of scalar-to-vector identities in the chapter for scalar-to-matrix derivatives.
- Let $J = \|V\|_F^2$. Show that $\frac{\partial J}{\partial V} = 2V$. You may find it helpful to express the Frobenius norm as the sum of vector norms and then use scalar-to-vector identities.
- Let $J = \|AV\|_F^2$. Express J using vector norms and the columns of V . Show that $\frac{\partial J}{\partial V} = 2A^T AV$ by using the scalar-to-vector identities discussed in the chapter. Now show that the derivative of $J = \|AV + B\|^2$ is $2A^T(AV + B)$, where B is an $n \times k$ matrix. What you just derived is useful for gradient descent in matrix factorization.

The full solution is given in Chapter 8, where the derivative is computed using matrix calculus.

- 19.** Consider an additively separable multivariate function of the form $J(w_1, w_2, \dots, w_{100}) = \sum_{i=1}^{100} J_i(w_i)$. Each $J_i(w_i)$ is a univariate function, which has one global optimum and one local optimum. Discuss why the chances of coordinate descent to reach the global optimum with a randomly chosen starting point are likely to be extremely low.

This problem has 2^{100} local optima, by using the combinations of the local optima on different dimensions. Furthermore, once coordinate descent reaches a local optimum

on even one dimension, it cannot escape from it. Therefore, it is highly likely for coordinate descent to get stuck in one of the local optima.

- 20.** Propose a computational procedure to use single-variable coordinate descent in order to solve the L_2 -loss SVM. You may use line search for each univariate problem. Implement the procedure in a programming language of your choice.

This is an implementation exercise.

- 21.** Consider a bivariate quadratic loss function of the following form:

$$f(x, y) = ax^2 + by^2 + 2cxy + dx + ey + f$$

Show that $f(x, y)$ is convex if and only if a and b are non-negative, and c is at most equal to the geometric mean of a and b in absolute magnitude.

The Hessian of this function is a constant everywhere since it is quadratic, and is as follows:

$$H = \begin{pmatrix} 2a & 2c \\ 2c & 2b \end{pmatrix}$$

This matrix needs to be positive semi-definite. Therefore, the trace is non-negative, which means that $a + b \geq 0$. Also the determinant is non-negative, which means that $ab - c^2 \geq 0$. The latter also implies that $ab \geq 0$. Therefore, the sum and product of a and b are nonnegative, which means that a and b are non-negative. Furthermore, since $ab \geq c^2$, it means that c is at most equal to the geometric mean of a and b in absolute magnitude. It is possible for c to be negative.

- 22.** Show that the functions $f(\bar{x}) = \sqrt{\langle \bar{x}, \bar{x} \rangle}$ and $g(\bar{x}) = \langle \bar{x}, \bar{x} \rangle$ are both convex. With regard to inner products, you are allowed to use only the basic axioms, and the Cauchy-Schwarz/triangle inequality.

We first show that $f(\bar{x})$ is convex.

$$\begin{aligned} f(\lambda \bar{x} + (1 - \lambda) \bar{y}) &= \sqrt{\langle \lambda \bar{x} + (1 - \lambda) \bar{y}, \lambda \bar{x} + (1 - \lambda) \bar{y} \rangle} \\ &\leq \sqrt{\langle \lambda \bar{x}, \lambda \bar{x} \rangle} + \sqrt{\langle -(1 - \lambda) \bar{y}, -(1 - \lambda) \bar{y} \rangle} \quad [\text{Triangle Inequality}] \\ &= \sqrt{\lambda^2 \langle \bar{x}, \bar{x} \rangle} + \sqrt{(1 - \lambda)^2 \langle \bar{y}, \bar{y} \rangle} \quad [\text{Multiplicative axiom}] \\ &= \lambda \sqrt{\langle \bar{x}, \bar{x} \rangle} + (1 - \lambda) \sqrt{\langle \bar{y}, \bar{y} \rangle} \\ &= \lambda f(\bar{x}) + (1 - \lambda) f(\bar{y}) \end{aligned}$$

Therefore, the function $f(\bar{x})$ is convex. Furthermore, $g(\bar{x}) = f(\bar{x})^2$, where $f(\bar{x})$ is non-negative because of the positive definite axiom. Therefore, $g(\bar{x})$ is convex as well.

- 23. Two-sided matrix least-squares:** Let A be an $n \times m$ matrix and B be a $k \times d$ matrix. You want to find the $m \times k$ matrix X so that $J = \|C - AXB\|_F^2$ is minimized, where C is a known $n \times d$ matrix. Derive the derivative of J with respect to X and the optimality conditions. Show that one possible solution to the optimality conditions is $X = A^+CB^+$, where A^+ and B^+ represent the Moore-Penrose pseudo-inverses of A and B , respectively.

One can express the objective function by decomposing it into the individual terms:

$$J = \sum_i \sum_j (c_{ij} - \sum_k \sum_l a_{ik} x_{kl} b_{lj})^2$$

Therefore, we have the following:

$$\frac{\partial J}{\partial x_{pq}} = -2 \sum_i \sum_j a_{ip}(c_{ij} - \sum_k \sum_l a_{ik} x_{kl} b_{lj}) b_{qj} = -2 \bar{a}_p^T (C - AXB) \bar{b}_q^T$$

Here, \bar{a}_p is the p th column of A and \bar{b}_q is the q th row of B . One can write this in matrix calculus form:

$$\frac{\partial J}{\partial X} = -2A^T(C - AXB)B^T$$

Note that this is the derivative of J , and one can perform the optimization by performing the following updates after absorbing the factor of 2 in the learning rate:

$$X \leftarrow X + \alpha A^T(C - AXB)B^T$$

Alternatively, one may obtain the optimality conditions by setting the derivative to 0:

$$A^T(C - AXB)B^T = 0$$

Note that this system of equations may not have a unique solution. The unique solution exists if A is tall with linearly independent columns and B is wide with linearly independent rows. In such a case, the unique solution is shown in Exercise 49 of Chapter 2:

$$X = \underbrace{(A^T A)^{-1} A^T}_\text{Left Inverse} C \underbrace{B^T (B B^T)^{-1}}_\text{Right inverse}$$

However, in the general case, when nothing can be assumed on linear independence, *one possible solution* is $X^* = A^+ C B^+$, and there might be alternative solutions. The fact that X^* is one possible solution can be proven by showing that this solution satisfies the optimality conditions:

$$\begin{aligned} A^T(C - AX^*B)B^T &= A^T C B^T - A^T A X^* B B^T \\ &= A^T C B^T - \underbrace{A^T A A^+}_{A^T} C \underbrace{B^+ B B^T}_{B^T} \\ &= A^T C B^T - A^T C B^T = 0 \end{aligned}$$

Note that we used the result of Exercise 51 of Chapter 2 in order to simplify the above equation.

- 24.** Suppose that you replace the sum-of-squared-Euclidean objective with a sum-of-Manhattan objective for the k -means algorithm. Show that block coordinate descent results in the k -medians clustering algorithm, where the each dimension of the “centroid” representative is chosen as the median of the cluster along that dimension and assignment of points to representatives is done using the Manhattan distance instead of Euclidean distance.

The basic strategy for block coordinate descent is the same, where the assignment variables define one block and the representative variables form the other block. The argument for this is that the assignment step does not change by the change in objective function, since a point is always assigned to its closest representative. The only difference is that Manhattan distance is used to choose closest representatives

for for assignment of points to representatives, whereas the k -means algorithm uses the Euclidean distance in order to choose closest representatives to data points. The determination of representatives is more fundamentally affected by the change in the objective function. In each cluster, the representative minimizing the Manhattan distance is the dimension-wise median.

25. Consider the cubic polynomial objective function $f(x) = ax^3 + bx^2 + cx + d$. Under what conditions does this objective function not have a critical point? Under what conditions is it strictly increasing in $[-\infty, +\infty]$?

The derivative of this expression is $3ax^2 + 2bx + c$. This quadratic does not have a zero when its discriminant is negative. This occurs when $4b^2 - 12ac < 0$. In other words, we must have $b^2 - 3ac < 0$. The function is strictly increasing only when this condition is true and we have $3ax^2 + 2bx + c > 0$. This occurs when $a > 0$.

26. Consider the cubic polynomial objective function $f(x) = ax^3 + bx^2 + cx + d$. Under what conditions does this objective have exactly one critical point? What kind of critical point is it? Give an example of such an objective function.

By a similar argument as the previous exercise the derivative must have exactly one root. Therefore, we must have $b^2 - 3ac = 0$. This type of critical point is a saddle point because the sign of the derivative is the same on both sides of the critical point. An example of such a function is $f(x) = x^3$.

27. Let $f(x)$ be a univariate polynomial of degree n . What is the maximum number of critical points of this polynomial? What is the maximum number of minima, maxima, and saddle points?

The derivative of $f(x)$ is a polynomial of degree $(n - 1)$, which has at most $(n - 1)$ roots according to the fundamental theorem of algebra. Of these, only alternate roots can be minima or maxima. Therefore, at most $\lceil (n - 1)/2 \rceil = \lfloor n/2 \rfloor$ roots are minima or maxima. Furthermore, a saddle point requires a repeated root, of the degree $(n - 1)$ polynomial. Therefore, there are at most $\lfloor (n - 1)/2 \rfloor = \lceil n/2 \rceil - 1$ saddle points.

28. What is the maximum number of critical points of a polynomial of degree n in d dimensions? Give an example of a polynomial where this maximum is met.

The maximum number of critical points is $(n - 1)^d$. This maximum is met for linearly separable functions in which each univariate function has $(n - 1)$ critical points.

29. Suppose that \bar{h} and \bar{x} are column vectors, and W_1 , W_2 , and W_3 are matrices satisfying $\bar{h} = W_1W_2\bar{x} - W_2^2W_3\bar{x} + W_1W_2W_3\bar{x}$. Derive an expression for $\frac{\partial \bar{h}}{\partial \bar{x}}$.

In this case, one can use the relationship that if $\bar{h} = C\bar{x}$, then the vector-to-vector derivative of \bar{h} with respect to \bar{x} is C^T . In this case, we can set $C = W_1W_2 - W_2^2W_3 + W_1W_2W_3$ in order to obtain the final expression as $W_2^TW_1^T - W_3^TW_2^TW_2^T + W_3^TW_2^TW_1^T$.

30. Consider a situation in which $\bar{h}_i = W_iW_{i-1}\bar{h}_{i-1}$, for $i \in \{1 \dots n\}$. Here, each W_i is a matrix and each \bar{h}_i is a vector. Use the vector-centric chain rule to derive an expression for $\frac{\partial \bar{h}_i}{\partial \bar{h}_0}$.

It is easy to show that $\frac{\partial \bar{h}_i}{\partial \bar{h}_{i-1}} = W_{i-1}^TW_i^T$. Then, by using the vector-centric chain rule, we obtain the expression $W_0^T(\prod_{i=1}^{n-1} W_i^TW_i^T)W_n^T$.

Chapter 5

Optimization Challenges and Advanced Solutions

1. Consider the loss function $L = x^2 + y^{10}$. Implement a simple steepest-descent algorithm to plot the coordinates as they vary from the initialization point to the optimal value of 0. Consider two different initialization points of (0.5, 0.5) and (2, 2) and plot the trajectories in the two cases at a constant learning rate. What do you observe about the behavior of the algorithm in the two cases?

This is an implementation exercise.

2. As shown in this chapter, the number of steps taken by gradient descent is very sensitive to the scaling of the variables. In this exercise, we will show that the Newton method is completely insensitive to the scaling of the variables. Let \bar{x} be the set of optimization variables for a particular optimization problem (OP). Suppose we transform \bar{x} to \bar{y} by the linear scaling $\bar{y} = B\bar{x}$ with invertible matrix B , and pose the same optimization problem in terms of \bar{y} . The objective function might be non-quadratic. Show that the sequences $\bar{x}_0, \bar{x}_1 \dots \bar{x}_r$ and $\bar{y}_0, \bar{y}_1 \dots \bar{y}_r$ obtained by iteratively applying Newton's method will be related as follows:

$$\bar{y}_k = B\bar{x}_k \quad \forall k \in \{1 \dots r\}$$

A function $f(\bar{x})$ becomes $f(B^{-1}\bar{y})$. The gradient of the objective function is $B^{-1}\nabla f(B^{-1}\bar{y})$. The Hessian is $B^{-1}H(B^{-1}\bar{y})B^{-1T}$. Let us assume that $\bar{x}_i = \bar{y}_i$ for $i = 1 \dots t$. On substituting these values in the Newton update, we obtain the desired result.

3. Write down the second-order Taylor expansion of each of the following functions about $x = 0$: (a) x^2 ; (b) x^3 ; (c) x^4 ; (d) $\cos(x)$.
(a) x^2 , (b) 0, (c) 0, (d) $1 - x^2/2$
4. Suppose that you have the quadratic function $f(x) = ax^2 + bx + c$ with $a > 0$. It is well known that this quadratic function takes on its minimum value at $x = -b/2a$. Show that a single Newton step starting at any point $x = x_0$ will always lead to $x = -b/2a$ irrespective of the starting point x_0 .

The gradient is $2ax_0 + b$ and the second derivative is $2a$. The Newton update is:

$$x \leftarrow x_0 - (2ax_0 + b)/2a = -b/2a$$

5. Consider the objective function $f(x) = [x(x-2)]^2 + x^2$. Write the Newton update for this objective function starting at $x = 1$.

The first derivative at $x = 1$ is 2. The second derivative is $12x^2 - 24x + 8 + 2 = -2$. The Newton update is therefore $x \leftarrow x - 2/(-2) = x + 1 = 2$. Note that The Newton update increases the objective function value, and this objective function is not convex.

6. Consider the objective function $f(x) = \sum_{i=1}^4 x^i$. Write the Newton update starting at $x = 1$.

The first derivative is $4 + 3 + 2 + 1 = 10$. The second derivative is $12 + 6 + 2 = 20$. The update is $x \leftarrow x - 10/20 = 1 - 0.5 = 0.5$.

7. Is it possible for a Newton update to reach a maximum rather than a minimum? Explain your answer. In what types of functions is the Newton method guaranteed to reach either a maximum or a minimum?

The Newton update can reach any type of critical point. For concave functions, the Newton method will always be able to reach a maximum.

8. Consider the objective function $f(x) = \sin(x) - \cos(x)$, where the angle x is measured in radians. Write the Newton update starting at $x = \pi/8$.

The first derivative is $\cos(\pi/8) + \sin(\pi/8) = 0.924 + 0.383 = 1.307$. The second derivative is $\cos(\pi/8) - \sin(\pi/8) = 0.924 - 0.383 = 0.541$. The update is $x \leftarrow x - 1.307/0.541 = \pi/8 - 1.307/0.541 = -2.02$.

9. The Hessian H of a strongly convex quadratic function always satisfies $\bar{x}^T H \bar{x} > 0$ for any nonzero vector \bar{x} . For such problems, show that all conjugate directions are linearly independent.

Suppose that the conjugate directions are not linearly independent. Without loss of generality, assume that \bar{q}_d can be expressed in terms of $\bar{q}_1 \dots \bar{q}_{d-1}$.

$$\bar{q}_d = \sum_{i=1}^{d-1} \lambda_i \bar{q}_i \quad (5.1)$$

Then, we have $\bar{q}_d^T H \bar{q}_d = 0$ by expanding one of the two \bar{q}_d in the above expression in terms of its linear dependents. However, this is in contradiction to the fact that H is positive definite.

10. Show that if the dot product of a d -dimensional vector \bar{v} with d linearly independent vectors is 0, then \bar{v} must be the zero vector.

Let $\bar{x}_1 \dots \bar{x}_d$ be the d linearly independent vectors. These vectors form a basis for d -dimensional space, and therefore the vector \bar{v} must be expressed as a linear combination of $\bar{x}_1 \dots \bar{x}_d$. Therefore, we have $\bar{v} = \sum_{i=1}^d \alpha_i \bar{x}_i$. By taking the dot product of both sides with \bar{v} we get the following:

$$\|\bar{v}\|^2 = \sum_{i=1}^d \alpha_i (\bar{v} \cdot \bar{x}_i) = \sum_{i=1}^d \alpha_i (0) = 0 \quad (5.2)$$

The modulus of a vector is zero, when the vector is itself zero.

11. The chapter uses steepest descent directions to iteratively generate conjugate directions. Suppose we pick d arbitrary directions $\bar{v}_0 \dots \bar{v}_{d-1}$ that are linearly independent. Show that (with appropriate choice of β_{ti}) we can start with $\bar{q}_0 = \bar{v}_0$ and generate successive conjugate directions in the following form:

$$\bar{q}_{t+1} = \bar{v}_{t+1} + \sum_{i=0}^t \beta_{ti} \bar{q}_i \quad (5.3)$$

Discuss why this approach is more expensive than the one discussed in the chapter.

Pre-multiply both sides of Equation 5.3 with the row vector $\bar{q}_i^T H$ and use the conjugacy condition to set the LHS to 0. This results in the following value of β_{ti} :

$$\beta_{ti} = -\frac{\bar{q}_i^T H \bar{v}_{t+1}}{\bar{q}_i^T H \bar{q}_i} \quad (5.4)$$

This approach requires us to maintain *all* previous directions, and also requires us to compute $O(d)$ values of β_{ti} for varying i . Therefore, the approach is not as time- and space-efficient.

12. The definition of β_t ensures that \bar{q}_t is conjugate to \bar{q}_{t+1} . This exercise systematically shows that any direction \bar{q}_i for $i \leq t$ satisfies $\bar{q}_i^T H \bar{q}_{t+1} = 0$.

- (a) Recall that $H\bar{q}_i = [\nabla L(\bar{W}_{i+1}) - \nabla L(\bar{W}_i)]/\delta_i$ for quadratic loss functions, where δ_i depends on i th step-size. Show the following for all $i \leq t$:

$$\delta_i [\bar{q}_i^T H \bar{q}_{t+1}] = -[\nabla L(\bar{W}_{i+1}) - \nabla L(\bar{W}_i)]^T [\nabla L(\bar{W}_{t+1})] + \delta_i \beta_t (\bar{q}_i^T H \bar{q}_t)$$

Also show that $[\nabla L(\bar{W}_{t+1}) - \nabla L(\bar{W}_t)] \cdot \bar{q}_i = \delta_t \bar{q}_i^T H \bar{q}_t$.

- (b) Show that $\nabla L(\bar{W}_{t+1})$ is orthogonal to each \bar{q}_i for $i \leq t$. [The proof for the case when $i = t$ is trivial because the gradient at line-search termination is always orthogonal to the search direction.]
- (c) Show that the loss gradients at $\bar{W}_0 \dots \bar{W}_{t+1}$ are mutually orthogonal.
- (d) Show that $\bar{q}_i^T H \bar{q}_{t+1} = 0$ for $i \leq t$. [The case for $i = t$ is trivial.]

Since $H\bar{q}_i = [\nabla L(\bar{W}_{i+1}) - \nabla L(\bar{W}_i)]/\delta_i$, we can use the transpose of the vector condition while keeping in mind that H is symmetric:

$$\bar{q}_i^T H = [\nabla L(\bar{W}_{i+1}) - \nabla L(\bar{W}_i)]^T / \delta_i \quad (5.5)$$

Now note that successive conjugate directions are generated as $\bar{q}_{t+1} = -\nabla L(\bar{W}_{t+1}) + \beta_t \bar{q}_t$. Pre-multiplying both sides with $\bar{q}_i^T H$ for $i \leq t$, we get the following:

$$\bar{q}_i^T H \bar{q}_{t+1} = -\bar{q}_i^T H [\nabla L(\bar{W}_{t+1})] + \beta_t \bar{q}_i^T H \bar{q}_t \quad (5.6)$$

Now substituting for only the first $\bar{q}_i^T H$ in the right-hand side of Equation 5.6 using Equation 5.5, we get the desired result.

Also since we have $H\bar{q}_t = [\nabla L(\bar{W}_{t+1}) - \nabla L(\bar{W}_t)]/\delta_t$, we can pre-multiply both sides with the row vector \bar{q}_i^T to get the second result of (a). Note that the right-hand side can also be written as a dot product, since it is the product of a row vector and a column vector.

Joint proof of (b), (c), (d): We will make the inductive assumption that all the statements of (b), (c), and (d) are true for values of t that are less than or equal to $k - 1$, and we will show the results at $t = k$.

First, from (a), we have already proved that $\nabla L(\bar{W}_{k+1}) \cdot \bar{q}_i = \nabla L(\bar{W}_k) \cdot \bar{q}_i + \delta_k \bar{q}_i^T H \bar{q}_k$ for $i < k$. Both the terms on the right-hand side are zero because of the inductive assumption in the case where $i < k - 1$. For the case when $i = k - 1$ both terms are also zero because of the line-search condition $\nabla L(\bar{W}_k) \cdot \bar{q}_{k-1} = 0$ and also because of the fact that $\bar{q}_{k-1}^T H \bar{q}_k = 0$ by definition. Therefore, we have shown the induction for (b).

Next, we will examine the orthogonality of gradient $\nabla L(\bar{W}_{k+1})$ with gradient of $\nabla L(\bar{W}_i)$. In the case where $i = k$, we know that $\nabla L(\bar{W}_{k+1}) \cdot \bar{q}_k = 0$ because of the line-search condition. Now expanding $\bar{q}_k = \beta_{k-1} \bar{q}_{k-1} - \nabla L(\bar{W}_k)$ and taking the dot product of both sides with $\nabla L(\bar{W}_{k+1})$, we get $0 = \nabla L(\bar{W}_{k+1}) \cdot [\beta_{k-1} \bar{q}_{k-1} - \nabla L(\bar{W}_k)]$. Note that we have used the line-search condition to set the LHS to 0. Now note that the first of the two terms on the RHS sets to 0 because of the inductive assumption, which leaves only the term involving successive gradients. Therefore, we get the result that *immediately successive gradients* $\nabla L(\bar{W}_k)$ and $\nabla L(\bar{W}_{k+1})$ are orthogonal.

Next, we will show the result that $\nabla L(\bar{W}_i)$ is orthogonal to $\nabla L(\bar{W}_{k+1})$ for $i < k$. By rearranging the recursive definition of \bar{q}_i in terms of gradient at \bar{W}_i and \bar{q}_{i-1} , the gradient $\nabla L(\bar{W}_i)$ can be expressed as a linear combination of \bar{q}_i and \bar{q}_{i-1} . Both \bar{q}_i and \bar{q}_{i-1} are orthogonal to $\nabla L(\bar{W}_{k+1})$ based on the inductive assumption, and therefore any linear combination of them must also be orthogonal to $\nabla L(\bar{W}_{k+1})$. In other words, $\nabla L(\bar{W}_i)$ is orthogonal to $\nabla L(\bar{W}_{k+1})$ for $i < k$. Therefore, we have shown the induction for (c).

Next, we will show the induction for (d) only for the case when $i < k$. This is because conjugacy between immediately successive directions follows from the way in which \bar{q}_{k+1} is defined as a function of \bar{q}_k and how β_k is chosen. We restate the result we showed in (a) using the inductive index k .

$$\delta_i [\bar{q}_i^T H \bar{q}_{k+1}] = -[\nabla L(\bar{W}_{i+1}) - \nabla L(\bar{W}_i)]^T [\nabla L(\bar{W}_{k+1})] + \delta_i \beta_k (\bar{q}_i^T H \bar{q}_k)$$

Now observe that the first term on the RHS is 0 because of what we proved in (c). The second term on the RHS is 0 because of the inductive assumption. Therefore, we have shown that the LHS is 0 as well. This completes the induction for (d), which is the conjugacy condition.

The initialization conditions of the induction can be shown in a relatively simple way by using the fact that the gradient at the optimal point found by line search is always orthogonal to the original direction, and also the fact that immediately successive directions are always conjugate (by definition).

13. Consider a setting in which your data set has a smaller number of points than the number of dimensions, and you are using the Newton method in conjunction with a regularized L_2 -loss SVM. Discuss how you can use this fact to make the update more efficient.

Note that the solution $(D^T \Delta_w D + \lambda I_d)^{-1} D^T \Delta_w \bar{y}$ can also be written as $(D^T \Delta_w D + \lambda I_d)^{-1} D^T \sqrt{\Delta_w} \bar{y}$ because of the binary nature of the diagonal matrix. Set $D_w = \sqrt{\Delta_w} D$. Then, the above matrix is $(D_w^T D_w + \lambda I_d)^{-1} D_w^T \bar{y}$, which is equivalent to

$D_w^T(D_w D_w^T + \lambda I_n)^{-1} \bar{y}$ by the Woodbury identity. The latter inverts a smaller matrix, since n is much less than d .

- 14. Saddle points are ubiquitous in high dimensions:** Consider the function $f(x) = x^3 - 3x$ with a minimum at $x = -1$ and a maximum at $x = 1$. Define the following multivariate function:

$$F(x_1 \dots x_d) = \sum_{i=1}^d f(x_i)$$

Show that this function has one minimum, one maximum, and $2^d - 2$ saddle points. Argue why high-dimensional functions have proliferating saddle points.

The single minimum is a vector containing only the values of -1 , and the single maximum is a vector containing only the values of $+1$. Any combination of $+1$ and -1 for the variables is a saddle point. High-dimensional functions have proliferating saddle points because all directions must be either increasing or decreasing at a given critical point for it to be a minimum or a maximum. This is exponentially less likely than the different directions to be a combination of increasing or decreasing directions.

- 15.** Give a proof of the unified Newton update for machine learning.

The first step is to show that the first gradient is of the form $D^T \Delta_1 \bar{I} + \lambda \bar{W}$. Note that the point specific gradient for the loss term is given by $L'(y, z_i) \bar{X}_i^T$ by the use of the chain rule. After adding the point specific gradients and the regularization gradient, we get the following overall gradient:

$$\nabla J = \lambda \bar{W} + \sum_{i=1}^n L'(y, z_i) \bar{X}_i^T = D^T \Delta_1 \bar{I} + \lambda \bar{W}$$

Similarly, the point specific Hessian of the loss component can be shown to be $L''(y_i, z_i) \bar{X}_i \bar{X}_i^T$. After adding the point-specific Hessians and the effect of regularization, we get the following overall Hessian:

$$H = D^T \Delta_2 D + \lambda I$$

Therefore, by combining the Hessians and gradient we obtain the overall Newton update as follows:

$$\bar{W} \leftarrow \bar{W} - \alpha H^{-1} \nabla J$$

The result follows.

- 16.** Consider a directed-acyclic graph G (i.e., graph without cycles) with source node s and sink t . Each edge is associated with a length and a multiplier. The length of a path from s to t is equal to the sum of the edge lengths on the path and the multiplier of the path is the product of the corresponding edge multipliers. Devise dynamic programming algorithms to find (i) the longest path from s to t , (ii) the shortest path from s to t , (iii) the average path length from s to t , and (iv) the sum of the path-multipliers of all paths from s to t .

In each case, nodes are visited starting from the source. A node is visited only when all its incoming nodes have been visited. If all incoming nodes have been visited for multiple nodes then ties are broken arbitrarily. Each visited node maintains the value of interest. When the sink is reached, its value of interest is automatically computed.

- (i) Each visited node maintains its longest path length, with the source being initialized to 0. For a visited node we set the longest path length of the node as the longest of all possibilities among all incoming edges by adding incoming edge length to the length label on incoming node.
 - (ii) This is the same as (i), except that we select the shortest path rather than the longest path.
 - (iii) We maintain two separate values with nodes corresponding to the number of paths and the sum of path lengths. The sum of path lengths at a node is equal to the sum of incoming edge length and all the path length values at incoming nodes. The source is initialized to 1. The number of paths at a node is equal to sum of the number of paths at incoming nodes, with the source being initialized to 1. At the sink, we divide the sum of path lengths with the number of paths to yield the average path length.
 - (iv) We maintain the sum of path multipliers at the nodes. The value at a visited node is obtained by multiplying each incoming edge length with the value on the corresponding incoming node and adding all these values.
- 17.** *Give an example of a univariate cubic along with two possible starting points for Newton's method, which terminate in maxima and minima, respectively.*

Consider the following cubic function:

$$f(x) = x^3 - 6x^2 + 9x - 10$$

It is easy to set the derivative to 0 and find that the critical points are $x = 1$ and $x = 3$. The former is a maximum, whereas the latter is a minimum. Furthermore, the second derivative is $6x - 12$, which is negative for $x > 2$ and positive for $x < 2$. It can be shown that starting the Newton method at any point $x < 2$, where the 1×1 Hessian is positive definite will converge to the minimum solution. On the other hand, starting the Newton method at $x > 2$, where the Hessian is negative definite, will converge to the maximum solution.

- 18.** *Linear regression with L_1 -loss minimizes $\|D\bar{W} - \bar{y}\|_1$ for data matrix D and target vector \bar{y} . Discuss why the Newton method cannot be used in this case.*

The function is not twice differentiable. Therefore, the Hessian cannot be computed. This is a similar reason to why the L_1 -loss SVM cannot be used in this case.

Chapter 6

Lagrangian Relaxation and Duality

1. Suppose you want to find the largest area of rectangle that can be inscribed in a circle of radius 1. Formulate a 2-variable optimization problem with constraints to solve this problem. Discuss how you can convert this problem into a single-variable optimization problem without constraints.

Let x be length of the rectangle and y be its width. The diagonal of the rectangle is the diameter of the circle, and therefore we have $x^2 + y^2 = 4$, which is the constraint. The objective function is xy . One can eliminate y in order to create the area $x\sqrt{4 - x^2}$.

2. Consider the following optimization problem:

$$\begin{aligned} &\text{Minimize } x^2 + 2x + y^2 + 3y \\ &\text{subject to:} \\ &x + y = 1 \end{aligned}$$

Suppose that (x_0, y_0) is a point satisfying the constraint $x + y = 1$. Compute the projected gradient at (x_0, y_0) .

The unconstrained gradient at (x_0, y_0) is $[2x_0 + 2, 2y_0 + 3]^T$. Here, the matrix A is $[1, 1]$. Therefore, we need to compute the matrix $(I - A^T(AA^T)^{-1}A)$. One can show that this is a 2×2 matrix with 0.5 on each diagonal and -0.5 on each off-diagonal. Pre-multiplying the unconstrained gradient with this matrix, we obtain the fact that the projected gradient is $[2x_0 - 2y_0 - 1, -2x_0 + 2y_0 + 1]^T$. Note that the gradient is such that increasing x by a particular amount with decrease y by the same amount, and vice versa.

3. Use the method of variable transformation to eliminate both the constraint and variable y in Exercise 2. Compute the optimal solution of the resulting unconstrained problem. What is the optimal objective function value?

We simply substitute $y = 1 - x$ in order to obtain a modified objective function, which is $x^2 + 2x + (1 - x)^2 + 3(1 - x) = 2x^2 - 3x + 4$. This function takes on its minimum at $x = 3/4$. The optimum objective function value is $23/4$.

4. *Compute the dual of the objective function in Exercise 2. Compute the optimal solution as well as the resulting objective function value.*

The dual of the objective function is given by $x^2 + 2x + y^2 + 3y + \alpha(x + y - 1)$. The gradient with respect to x results in $2x + 2 + \alpha = 0$ and the gradient with respect to y results in $2y + 3 + \alpha = 0$. Eliminating α , we obtain $2(x - y) = 1$. Combining with the condition $x + y = 1$, we obtain $4x = 3$ or $x = 3/4$. The value of α is $-7/2$, and y is $1/4$. Note that the dual has a single feasible point in this case. The dual objective function can be written as follows:

$$L(\alpha) = \alpha^2/2 - 13/4$$

This relaxation has a single feasible point at $\alpha = -7/2$. The optimal dual objective function value for this single feasible point is $49/8 - 13/4$ which is equal to $23/8$. One can easily verify that the optimal primal solution also evaluates to the same value.

5. *Implement a gradient-descent algorithm for linear regression with box constraints. Use Python or any other programming language of your choice.*

This is an implementation algorithm.

6. **Linear programming dual:** *Consider the following linear programming optimization problem with respect to primal variables $\bar{w} = [w_1, w_2, \dots, w_d]^T$:*

$$\begin{aligned} & \text{Minimize } \sum_{i=1}^d c_i w_i \\ & \text{subject to:} \\ & A\bar{w} \leq \bar{b} \end{aligned}$$

Here, A is an $n \times d$ matrix, and \bar{b} is an n -dimensional column vector. Formulate the dual of this optimization problem by using the Lagrangian relaxation. Are there any conditions under which strong duality holds?

Let $\bar{\alpha}$ be the column vector of Lagrange multipliers. Then, the Lagrangian relaxation is $\sum_{i=1}^d c_i w_i + \bar{\alpha}^T (A\bar{w} - \bar{b})$. Differentiating with respect to \bar{w} and using matrix calculus, we obtain $A^T \bar{\alpha} + \bar{c} = 0$. This is the primal dual-constraint, and in addition we have $\bar{\alpha} \geq 0$. One can also eliminate \bar{w} in the objective function setting $\bar{c} = A^T \bar{\alpha}$, and simply obtaining the objective function $-\bar{b}^T \bar{\alpha}$. Therefore, the linear programming dual is that of minimizing $\bar{b}^T \bar{\alpha}$ subject to $A^T \bar{\alpha} + \bar{c} = 0$ and $\bar{\alpha} \geq 0$. Strong duality always holds for this relaxation.

7. **Quadratic programming dual:** *Consider the following quadratic programming optimization problem with respect to primal variables $\bar{w} = [w_1, w_2, \dots, w_d]^T$*

$$\begin{aligned} & \text{Minimize } \frac{1}{2} \bar{w}^T Q \bar{w} + \sum_{i=1}^d c_i w_i \\ & \text{subject to:} \\ & A\bar{w} \leq \bar{b} \end{aligned}$$

Here, Q is a $d \times d$ matrix, A is an $n \times d$ matrix, and \bar{b} is an n -dimensional column vector. Formulate the dual of this optimization problem by using the Lagrangian relaxation. Assume that Q is invertible. Are there any conditions under which strong duality holds?

In this case, the Lagrangian relaxation is $\bar{w}^T Q \bar{w} / 2 + \sum_{i=1}^d c_i w_i + \bar{\alpha}^T (A \bar{w} - \bar{b})$. Computing the gradient with respect to \bar{w} we obtain $Q \bar{w} + \bar{c} + A^T \bar{\alpha} = 0$. Therefore, $\alpha^T A$ in the objective function can be replaced with $-(Q \bar{w} + \bar{c})^T$. Furthermore, we also have $\bar{w} = -Q^{-1}[\bar{c} + A^T \bar{\alpha}]$. Therefore, the objective function is the following:

$$\begin{aligned} J &= \bar{w}^T Q \bar{w} / 2 + \sum_{i=1}^d c_i w_i + \bar{\alpha}^T (A \bar{w} - \bar{b}) \\ &= \bar{w}^T Q \bar{w} / 2 + \bar{c}^T \bar{w} - (\bar{w}^T Q + \bar{c}^T) \bar{w} - \bar{\alpha}^T \bar{b} \\ &= -\bar{w}^T Q \bar{w} / 2 - \bar{\alpha}^T \bar{b} \\ &= -(\bar{c} + A^T \bar{\alpha})^T Q^{-1} (\bar{c} + A^T \bar{\alpha}) / 2 - \bar{\alpha}^T \bar{b} \end{aligned}$$

Therefore, one wants to minimize $(\bar{c} + A^T \bar{\alpha})^T Q^{-1} (\bar{c} + A^T \bar{\alpha}) / 2 + \bar{\alpha}^T \bar{b}$ subject to the constraint that $\bar{\alpha}$ is nonnegative.

8. Consider the SVM optimization problem where we explicitly allow a bias variable b . In other words, the primal SVM optimization problem is stated as follows:

$$J = \sum_{i=1}^n \max\{0, (1 - y_i [\bar{W} \cdot \bar{X}_i^T] + b)\} + \frac{\lambda}{2} \|\bar{W}\|^2$$

Compute the dual of this optimization formulation by using analogous steps to those discussed in the chapter. How would you handle the additional constraint in the dual formulation during gradient descent?

The dual objective function turns out to be exactly the same, except that we have the equality constraint $\sum_i \lambda_i y_i = 0$. For gradient descent, we can use the ideas for gradient-descent with equality constraints.

9. The primal formulation for least-squares regression can be recast in terms of similarities s_{ij} between pairs of data points as follows:

$$J = \frac{1}{2} \sum_{i=1}^n (y_i - \sum_{p=1}^n \beta_p s_{pi})^2 + \frac{\lambda}{2} \sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j s_{ij}$$

Here, s_{ij} is the similarity between points i and j . Convert this unconstrained optimization problem into a constrained problem, and evaluate the dual of the problem in terms of s_{ij} .

10. Let $\bar{z} \in \mathcal{R}^d$ lie outside the ellipsoid $\bar{x}^T A \bar{x} + \bar{b}^T \bar{x} + c \leq 0$, where A is a $d \times d$ positive semi-definite matrix and $\bar{x} \in \mathcal{R}^d$. We want to find the closest projection of \bar{z} on this convex ellipsoid to enable projected gradient descent. Use Lagrangian relaxation to show that the projection point \bar{z}_0 must satisfy the following:

$$\bar{z} - \bar{z}_0 \propto 2A\bar{z}_0 + \bar{b}$$

Interpret this condition geometrically in terms of the tangent to the ellipsoid.

The objective function is $\|\bar{z} - \bar{z}_0\|^2/2$, and the constraint $\bar{z}_0^T A \bar{z}_0 + \bar{b}^T \bar{z}_0 + c = 0$ is satisfied with equality since the point lies on the surface of the ellipsoid. The Lagrangian relaxation is as follows:

$$L(\bar{z}_0, \lambda) = \|\bar{z} - \bar{z}_0\|^2/2 - \lambda(\bar{z}_0^T A \bar{z}_0 + \bar{b}^T \bar{z}_0 + c)$$

There is no constraint of λ because the problem is equality constrained. Note that \bar{z} is treated as a constant vector. Setting the gradient with respect to \bar{z}_0 to 0, we obtain:

$$\bar{z} - \bar{z}_0 = \lambda(2A\bar{z}_0 + \bar{b})$$

Therefore, the result follows with λ as the proportionality constant.

Note that $\bar{z} - \bar{z}_0$ is the vector joining \bar{z} and \bar{z}_0 . The RHS is the normal to the tangent surface of the ellipsoid. In other words, the line joining \bar{z} and \bar{z}_0 must be perpendicular to the tangent surface of the ellipsoid.

11. Consider the following optimization problem:

$$\begin{aligned} & \text{Minimize } x^2 - y^2 - 2xy + z^2 \\ & \text{subject to:} \\ & x^2 + y^2 + z^2 \leq 2 \end{aligned}$$

Imagine that we are using coordinate descent in which we are currently optimizing the variable x , when y and z are set to 1 and 0, respectively. Solve for x . Then, solve for y by setting x and z to their current values. Finally, solve for z in the same way. Perform another full cycle of coordinate descent to confirm that coordinate descent cannot improve further. Provide an example of a solution with a better objective function value. Discuss why coordinate descent was unable to find an optimal solution.

In the first iteration, we try to minimize $x^2 - 2x$, subject to $x^2 \leq 1$, which yields the optimum at $x = 1$. In the next iteration, we obtain $y = 1$ as the optimum, and in the third we obtain $z = 0$. This is the final solution that does not change. However, the solutions $[0, \sqrt{2}, 0]$ and $[0.5, \sqrt{1.75}, 0]$ are two examples of better optima. This problem is not convex, and therefore, coordinate descent might not be successful.

12. Consider the dual objective function in Lagrangian relaxation, as a function of only the dual variables:

$$L(\bar{\alpha}) = \text{Minimize}_{\bar{w}} [F(\bar{w}) + \sum_{i=1}^m \alpha_i f_i(\bar{w})]$$

Show that $L(\bar{\alpha})$ is always concave in $\bar{\alpha}$, irrespective of the convexity structure of the original optimization problem.

$L(\bar{\alpha})$ is obtained as the minimum of an infinite number of linear functions in $\bar{\alpha}$ at fixed values of \bar{w} . A linear function is trivially concave. The minimum of a set of concave functions is concave even if there are infinite number of them.

13. **Nonnegative box regression:** Formulate the dual optimization problem for regularized linear regression with $n \times d$ data matrix D , regressand vector \bar{y} , and with nonnegativity constraints on the parameter vector.

Consider the linear regression problem of minimizing $\|D\bar{w} - \bar{y}\|^2/2 + \|\bar{w}\|^2/2$. One can use the solution of Exercise 7 after setting $Q = D^T D + \lambda I_d$, and $c_i = -[D^T \bar{y}]_i$. The constant term of $\bar{y}^T \bar{y}/2$ should be added to the dual objective function to ensure that the primal and dual have the same objective function value. Furthermore, the nonnegativity constraints can be expressed in the form $A\bar{w} \leq \bar{b}$ by setting $A = -I$ and $\bar{b} = \bar{0}$. Therefore, let us rewrite the dual objective function of Exercise 7 in maximization form:

$$\begin{aligned} J &= -(\bar{c} + A^T \bar{\alpha})^T Q^{-1} (\bar{c} + A^T \bar{\alpha})/2 - \bar{\alpha}^T \bar{b} \\ &= -(-D^T \bar{y} - \bar{\alpha})^T (D^T D + \lambda I_d)^{-1} (-D^T \bar{y} - \bar{\alpha})/2 + \bar{y}^T \bar{y}/2 \\ &= -(D^T \bar{y} + \bar{\alpha})^T (D^T D + \lambda I_d)^{-1} (D^T \bar{y} + \bar{\alpha}) + \bar{y}^T \bar{y}/2 \end{aligned}$$

One can also write this optimization function in minimization form:

$$J_{\min} = -J = (D^T \bar{y} + \bar{\alpha})^T (D^T D + \lambda I_d)^{-1} (D^T \bar{y} + \bar{\alpha}) - \bar{y}^T \bar{y}/2$$

In addition, we have the dual constraint that $\bar{\alpha} \geq 0$.

- 14. Hard Regularization:** Consider the case where instead of Tikhonov regularization, you solve the linear regression problem of minimizing $\|A\bar{x} - \bar{b}\|^2$ subject to the spherical constraint $\|\bar{x}\| \leq r$. Formulate the Lagrangian dual of the problem with variable $\alpha \geq 0$. Show that the dual variable plays the same role as the regularization parameter in Tikhonov regularization:

$$\bar{x} = (A^T A + \alpha I)^{-1} A^T \bar{b}$$

Under what conditions is α equal to 0? In the latter case, show that the optimal dual variable α is the solution to the secular equation:

$$\bar{b}^T A (A^T A + \alpha I)^{-2} A^T \bar{b} = r^2$$

This optimization problem becomes $\|A\bar{x} - \bar{b}\| + \alpha(\|\bar{w}\|^2 - r^2)$. On setting the gradient to zero one obtains $A^T(A\bar{x} - \bar{b}) + \alpha\bar{x} = 0$. This yields the same closed form as least-squares regression:

$$\bar{x} = (A^T A + \alpha I)^{-1} A^T \bar{b}$$

Note that the value of α is zero according to the complementary slackness conditions when the constraint is not satisfied tightly at equality. This happens when unconstrained least-squares regression already gives a solution whose norm is less than r^2 . On the other hand, if α is not zero, the bound $\|\bar{x}\|^2 = r^2$ is satisfied tightly. Substituting for $\bar{x} = (A^T A + \alpha I)^{-1} A^T \bar{b}$ in this equal, we obtain the secular equation.

- 15.** Propose a (primal) gradient-descent algorithm for the hard regularization model of the previous exercise. Use the projected gradient-descent method. The key point is in knowing how to perform the projection step.

One performs the same updates as unconstrained least-squares regression. However, the parameter vector is scaled to $\bar{w} \leftarrow r \bar{w} / \|\bar{w}\|$, whenever one moves outside the feasible region. This is the projection step.

- 16. Best subset selection:** Consider an $n \times d$ data matrix D in which you want to find the best subset of k features that are related to the n -dimensional regressand vector \bar{y} . Therefore, the following mixed integer program is formulated with d -dimensional real vector \bar{w} , d -dimensional binary vector \bar{z} , and an a priori (constant) upper bound M on each coefficient in \bar{w} . The optimization problem is to minimize $\|D\bar{w} - \bar{y}\|^2$ subject to the following constraints:

$$\bar{z} \in \{0, 1\}^d, \quad \bar{w} \leq M\bar{z}, \quad \bar{1}^T \bar{z} = k$$

The notation $\bar{1}$ denotes a d -dimensional vector of 1s. Propose an algorithm using block coordinate descent for this problem, where each optimized block contains just two integer variables and two real variables.

This algorithm is performed by using repeated interchange. We first select k features at random and set those values of z_i to 1. In each iteration, we pick pairs of z_i so that one of them is 1 and the other is 0. We test an interchange effect on objective function while keeping all other variables fixed. The resulting optimization problem is a box regression problem using k variables.

- 17. Duality Gap:** Suppose that you are running the dual gradient descent algorithm for the SVM, and you have the (possibly suboptimal) dual variables $\alpha_1 \dots \alpha_n$ in the current iteration. Propose a quick computational procedure to estimate an upper bound on how far this dual solution is from optimality.

The weight vector can be estimated from the dual variables as $\bar{W} = \sum_{i=1}^n \alpha_i y_i \bar{X}_i$. The primal solution can be used to compute the primal objective function value and the dual solution can be used to estimate the dual objective function value. The difference between the two is referred to as the duality gap, and provides an upper bound on the difference between the primal and dual solutions.

- 18.** State whether the following minimax functions $f(x, y)$ satisfy John von Neumann's strong duality condition, where x is the minimization variable and y is the maximization variable: (i) $f(x, y) = x^2 + 3xy - y^4$, (ii) $f(x, y) = x^2 + xy + y^2$, (iii) $f(x, y) = \sin(y - x)$, and (iv) $f(x, y) = \sin(y - x)$ for $0 \leq x \leq y \leq \pi/2$.

In order to satisfy strong duality, the function needs to be convex in the minimization variable and concave in the maximization variable. Therefore, we simply find the second derivative with respect to each of x and y in order to check for concavity or convexity. Using this approach we find that (i) and (iv) satisfy strong duality, whereas others do not.

- 19. Failure of coordinate descent:** Consider the problem of minimizing $x^2 + y^2$, subject to $x + y \geq 1$. Show using Lagrangian relaxation that the optimal solution is $x = y = 0.5$. Suppose that you start coordinate descent for this problem at $x = 1$ and $y = 0$. Discuss why coordinate descent will fail.

The Lagrangian relaxation is $x^2 + y^2 - \lambda(x + y - 1)$ and its gradient is $[2x - \lambda, 2y - \lambda]$. Setting the gradient to 0, we obtain $x = y = \lambda/2$. Therefore, we can eliminate y from the original optimization problem by substituting $y = x$ and recast it as that of minimizing $2x^2$ subject to $2x \geq 1$. The minimum value is achieved at $x = 0.5$. Therefore $y = x = 0.5$. Coordinate descent fails in this case because all feasible directions of movement worsen the objective function.

- 20.** *Propose a linear variable transformation for Exercise 19, so that coordinate descent will work on the reformulated problem.*

Consider the new set of variables $w_1 = x + y$ and $w_2 = x - y$. Then, the reformulated problem is to minimize $\frac{1}{2}(w_1^2 + w_2^2)$ subject to $w_1 \geq 1$. Using coordinate descent yields $w_1 = 1$ and $w_2 = 0$. Upon transforming back to the original variables, we obtain $x = y = 0.5$.

- 21.** *Formulate a variation of an SVM with hinge loss, in which the binary target (drawn from -1 or $+1$) is known to be nonnegatively correlated with each feature based on prior knowledge. Propose a variation of the gradient descent method by using only feasible directions.*

The optimization formulation is the same as the hinge-loss SVM, except that the weights are constrained to be nonnegative. One can use the same gradient descent approach except that any negative weight is reset to a zero.

Chapter 7

Singular Value Decomposition

1. Use SVD to show the push-through identity for any $n \times d$ matrix D :

$$(\lambda I_d + D^T D)^{-1} D^T = D^T (\lambda I_n + D D^T)^{-1}$$

We substitute the SVD of $D = Q \Sigma P^T$ on the left-hand side to show that it is equal to $P(\lambda I_d + \Sigma^T \Sigma)^{-1} P^T P \Sigma Q^T$. This is equal to $P(\lambda I_d + \Sigma^T \Sigma)^{-1} \Sigma Q^T$.

On making the same substitution, the right-hand side is equal to $P \Sigma^T Q^T (\lambda I_n + Q \Sigma \Sigma^T Q^T)^{-1}$. This can be shown to be equal to $P \Sigma^T (\lambda I_n + \Sigma \Sigma^T)^{-1} Q^T$.

In order to show that the above two results are equal, we need to show the following:

$$(\lambda I_d + \Sigma^T \Sigma)^{-1} \Sigma = \Sigma^T (\lambda I_n + \Sigma \Sigma^T)^{-1}$$

Here, the key point is that Σ is an $n \times d$ diagonal matrix. Furthermore, $\Sigma \Sigma^T$ and $\Sigma^T \Sigma$ are diagonal matrices with σ_i^2 on the diagonal. The difference is in terms of the sizes of the two matrices and the number of trailing zeros on the diagonal. The diagonal matrices help in showing the result. Both matrices can be shown to be diagonal $d \times n$ matrices which have $\sigma_{ii}/(\lambda + \sigma_{ii}^2)$ on the i th diagonal entry.

2. Let D be an $n \times d$ data matrix, and \bar{y} be an n -dimensional column vector containing the dependent variables of linear regression. The Tikhonov regularization solution to linear regression predicts the dependent variables of a test instance \bar{Z} : using the following equation:

$$\text{Prediction}(\bar{Z}) = \bar{Z} \bar{W} = \bar{Z} (D^T D + \lambda I)^{-1} D^T \bar{y}$$

Here, the vectors \bar{Z} and \bar{W} are treated as $1 \times d$ and $d \times$ matrices, respectively. Show using the result of Exercise 1, how you can write the above prediction purely in terms of similarities between training points or between \bar{Z} and training points.

One can use the result of the previous exercise to show that

$$\text{Prediction}(\bar{Z}) = \bar{Z} \bar{W} = \bar{Z} D (D D^T + \lambda I)^{-1} \bar{y}$$

One can write the above result as the product of $\bar{Z} D$ and $(D D^T + \lambda I)^{-1}$. Note that $\bar{Z} D$ is a row vector containing the dot product between the test and training instances and $D D^T$ contains similarities between pairs of training instances.

3. Suppose that you are given a truncated SVD $D \approx Q\Sigma P^T$ of rank- k . Show how you can use this solution to derive an alternative rank- k decomposition $Q'\Sigma'P'^T$ in which the unit columns of Q (or/and P) might not be mutually orthogonal and the truncation error is the same.

First, we express the decomposition as UV^T by absorbing Σ in Q . Then, we take any non-singular matrix Z of size $d \times d$ and express UV^T as $UZ^{-1}ZV^T$. Then we define $U' = UZ^{-1}$ and $V' = ZV^T$. It is easy to show that $UV^T = U'V'^T$. Subsequently, we can convert this two-way decomposition into a three-way decomposition by using the procedure described in the book.

4. Suppose that you are given a truncated SVD $D \approx Q\Sigma P^T$ of rank- k . Two of the non-zero singular values are identical. The corresponding right singular vectors are $[1, 0, 0]^T$ and $[0, 1, 0]^T$. Show how you can use this solution to derive an alternative rank- k SVD $Q'\Sigma'P'^T$ for which the truncation error is the same. At least some columns of matrices Q' and P' need to be non-trivially different from the corresponding columns in Q and P (i.e., the i th column of Q' should not be derivable from the i th column of Q by simply multiplying with either -1 or $+1$). Give a specific example of how you might manipulate the right singular vectors to obtain a non-trivially different solution.

The two columns containing these two singular vectors can be replaced with any orthonormal basis of these columns. Examples of such columns include $[1/\sqrt{2}, 1/\sqrt{2}, 0]^T$ and $[1/\sqrt{2}, -1/\sqrt{2}, 0]^T$. Once these right singular vectors have been found, the (changed) left singular vectors can be found using the formula $\bar{q} = D\bar{p}/\sigma$. Here, \bar{p} is a changed right-singular vector, σ is the singular value, and D is the data matrix.

5. Suppose that you are given a particular solution $\bar{x} = \bar{x}_0$ that satisfies the system of equations $A\bar{x} = \bar{b}$. Here, A is an $n \times d$ matrix, \bar{x} is a d -dimensional vector of variables, and \bar{b} is an n -dimensional vector of constants. Show that all possible solutions to this system of equations are of the form $\bar{x}_0 + \bar{v}$, where \bar{v} is any vector drawn from a vector space \mathcal{V} . Show that \mathcal{V} can be found easily using SVD. [Hint: Think about the system of equations $A\bar{x} = 0$.]

The vector \bar{v} must be drawn from the null space of A . Any linear combination of right singular vectors with zero singular values provides the null space of A .

6. Consider the $n \times d$ matrix D . Construct the $(n + d) \times (n + d)$ matrix B as follows:

$$B = \begin{bmatrix} 0 & D^T \\ D & 0 \end{bmatrix}$$

Note that the matrix B is square and symmetric. Show that diagonalizing B yields all the information needed for constructing the SVD of D . [Hint: Relate the eigenvectors of B to the singular vectors of SVD.]

If \bar{q} and \bar{p} are left and right singular vectors of D with singular value σ , we know that $D\bar{p} = \sigma\bar{q}$ and we know that $D^T\bar{q} = \sigma\bar{p}$. Using these facts, it is not difficult to show that $[\bar{p}^T, \bar{q}^T]^T$ is an eigenvector of B with eigenvalue σ . Therefore, by finding the $(d + n)$ -dimensional eigenvectors of B , we can chop them up into left and right singular vectors of D and use the eigenvalues of B as singular values of D . One issue is that $[\bar{p}^T, -\bar{q}^T]^T$ might also be an eigenvector of B , and such eigenvectors will not add to the information about singular vectors of D . In general any linear combination of this pair of eigenvectors will be an eigenvector. Therefore, each left and right singular vector pair in D will define a 2-dimensional eigenspace in B .

7. Consider the following matrix A whose SVD is given by the following:

$$A = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^T$$

Compute the inverse of A without explicitly materializing A .

The inverse of A is given by the following:

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/4 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$$

We are simply using the fact that the inverse of $Q\Sigma P^T$ is $P\Sigma^{-1}Q^T$.

8. Consider the following 2-way factorization of the matrix A :

$$A = UV^T = \begin{bmatrix} 4 & 1 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}^T$$

Convert this factorization into a 3-way factorization $Q\Sigma P^T$ in each of the following ways:

- (a) The L_2 -norm of each column of Q and P is 1.
- (a) The L_1 -norm of each column of Q and P is 1.

The second form of decomposition is used for nonnegative factorizations with probabilistic interpretability.

The two factorizations are as follows:

$$A = \begin{bmatrix} 4/5 & 1/\sqrt{5} \\ 3/5 & 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 5\sqrt{2} & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 2/\sqrt{5} \\ 1/\sqrt{2} & 1/\sqrt{5} \end{bmatrix}^T$$

$$A = \begin{bmatrix} 4/7 & 1/3 \\ 3/7 & 2/3 \end{bmatrix} \begin{bmatrix} 7*2 & 0 \\ 0 & 3*3 \end{bmatrix} \begin{bmatrix} 1/2 & 2/3 \\ 1/2 & 1/3 \end{bmatrix}^T$$

9. Suppose that you add a small amount of noise to each entry of an $n \times d$ matrix D with rank $r \ll d$ and $n \gg d$. The noise is drawn from a Gaussian distribution, whose variance $\lambda > 0$ is much smaller than the smallest non-zero singular value of D . The nonzero singular values of D are $\sigma_{11} \dots \sigma_{rr}$. What do you expect the rank of the modified matrix D' to become?

The rank of the modified matrix will become the smaller of n and d , which is d . In fact, the matrix formed by using any d rows of D (after perturbation) will have a nonzero determinant because the volume formed by the perturbed parallelepiped will be nonzero.

10. Consider the unconstrained optimization problem of minimizing the Frobenius norm $\|D - UV^T\|_F^2$, which is equivalent to SVD. Here, D is an $n \times d$ data matrix, U is an $n \times k$ matrix, and V is a $d \times k$ matrix.

- (a) Use differential calculus to show that the optimal solution satisfies the following conditions:

$$DV = UV^T V$$

$$D^T U = VU^T U$$

- (b) Let $E = D - UV^T$ be a matrix of errors from the current solutions U and V . Show that an alternative way to solve this optimization problem is by using the following gradient-descent updates:

$$\begin{aligned} U &\Leftarrow U + \alpha EV \\ V &\Leftarrow V + \alpha E^T U \end{aligned}$$

Here, $\alpha > 0$ is the step-size.

- (c) Will the resulting solution necessarily contain mutually orthogonal columns in U and V ?
- (a) The gradients with respect to U and V are $-DV + UV^T V$ and $-D^T U + VU^T V$. We are ignoring a factor of 2 for simplicity. Setting these gradients to zero, we obtain the desired result.
- (b) Note that we can use $E = D - UV^T$, and express the gradients as $-EV$ and $-E^T U$. Therefore, gradient-descent yields the above updates.
- (c) The resulting solution will not necessarily contain mutually orthogonal vectors, because an infinite number of alternate optima exist.
- 11.** Suppose that you change the objective function of SVD in Exercise 10 to add penalties on large values of the parameters. This is often done to reduce overfitting and improve generalization power of the solution. The new objective function to be minimized is as follows:

$$J = \|D - UV^T\|_F^2 + \lambda(\|U\|_F^2 + \|V\|_F^2)$$

Here, $\lambda > 0$ defines the penalty. How would your answers to Exercise 10 change?

- (a) The new gradients have λU and λV added to them (ignoring a factor of 2). Setting the new gradients to 0, we obtain:

$$\begin{aligned} DV &= UV^T V + \lambda U \\ D^T U &= VU^T U + \lambda V \end{aligned}$$

- (b) The updates are as follows:

$$\begin{aligned} U &\Leftarrow U(1 - \alpha\lambda) + \alpha EV \\ V &\Leftarrow V(1 - \alpha\lambda) + \alpha E^T U \end{aligned}$$

- (c) The columns are not necessarily mutually orthogonal.

- 12.** Recall from Chapter 3 that the determinant of a square matrix is equal to the product of its eigenvalues. Show that the determinant of a square matrix is also equal to the product of its singular values but only in absolute magnitude. Show that the Frobenius norm of the inverse of a $d \times d$ square matrix A is equal to the sum of squared inverses of the singular values of A .

The determinant of $Q\Sigma P^T$ is the product of the determinants of the three matrices. Orthogonal matrices have determinants of either 1 or -1, and the determinant of Σ is the product of the singular values. The result follows.

13. Show using SVD that a square matrix A is symmetric (i.e., $A = A^T$) if and only if $AA^T = A^T A$.

If $A = A^T$ then, $AA^T = A^T A = A^2$.

Conversely, if $AA^T = A^T A$, then the right singular vectors of A (eigenvectors of $A^T A$) are the same as the left singular vector vectors A (eigenvectors of AA^T). Furthermore, the corresponding eigenvalues are also the same. Therefore, the SVD of $A = Q\Sigma P^T$ satisfies $Q = P$. Any matrix of the form $P\Sigma P^T$ is symmetric.

14. Suppose that you are given the following valid SVD of a matrix:

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Is the SVD of this matrix unique? You may ignore multiplication of singular vectors by -1 as violating uniqueness. If the SVD is unique, discuss why this is the case. If the SVD is not unique, provide an alternative SVD of this matrix.

We will use the notations $D = Q\Sigma P^T$ to refer to the above matrices. The SVD is not unique, because two of the singular values are the same. One can change P to P_1 by using any orthonormal basis of the subspace defined by the last two columns of P . With the new matrix P_1 , the new matrix Q_1 can be reconstructed as $DP_1\Sigma^{-1}$. The resulting SVD is $Q_1\Sigma P_1^T$.

15. State a simple way to find the SVD of (a) a diagonal matrix with both positive and negative entries that are all different; and (b) an orthogonal matrix. Is the SVD unique in these cases?

(a) Q is set to the identity matrix, and then the k th column is multiplied with -1 , if the k th diagonal matrix is multiplied with -1 . Σ is set to the diagonal matrix and all its entries are flipped to positive sign. The matrix P is the identity matrix. This SVD is unique to the extent of multiplying columns of Q or P with -1 . (b) Q can be set to the given orthogonal matrix, and Σ, P are set to the identity matrix. Alternatively, Q, Σ can be set to the identity matrix, and P can be set to the transpose of the given matrix. Therefore, the SVD is obviously not unique.

16. Show that the largest singular value of $(A + B)$ is at most the sum of the largest singular values of each of A and B . Also show that the largest singular value of AB is at most the product of the largest singular values of A and B . Finally, show that the largest singular value of a matrix is a convex function of the matrix entries.

The largest singular value of a matrix is equal to the largest value of $\|A\bar{x}\|$ for any unit vector \bar{x} . This is the norm-constrained optimization problem. Therefore, for any unit vector \bar{x} , the value of $\|(A + B)\bar{x}\|$ is at most $\|A\bar{x}\| + \|B\bar{x}\|$ by the triangle inequality, which in turn is equal to the sum of the singular values. Note that multiplying a unique vector \bar{x} with AB scales up its length twice, once by B and then by A , and each scaling is the largest singular value of either matrix. Therefore, the product of the two matrices has a maximum singular value that is the corresponding product of their singular values. The maximum singular value is linear to scaling all entries of a matrix. In combination with the additive identity, it can be shown that the maximum singular value is convex.

17. If A is a square matrix, use SVD to show that AA^T and $A^T A$ are similar. What happens when A is rectangular?

Let $A = Q\Sigma P^T$. Here, Σ is square. We can show that $A^T A = P\Sigma^2 P^T$ and $AA^T = Q\Sigma^2 Q^T$. Since both have the same diagonal matrix, they are similar. This approach will not work for rectangular matrices since Σ will no longer be square, and the two matrices AA^T and $A^T A$ will have differing number of zero eigenvalues is $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$, respectively.

18. The Frobenius norm of a matrix A is defined as the trace of either AA^T or $A^T A$. Let P be a $d \times k$ matrix with orthonormal columns. Let D be an $n \times d$ data matrix. Use properties of the trace to show that the squared Frobenius norm of DP is the same as that of DPP^T . Interpret the matrices DP and DPP^T in terms of their relationship with D , when P contains the top- k right singular vectors of the SVD of D .

$$\|DPP^T\|_F^2 = \text{tr}(DPP^T(DPP^T)^T) = \text{tr}(DP \underbrace{(P^T P)}_I (DP)^T) = \text{tr}(DP(DP)^T) = \|DP\|_F^2$$

19. Consider two data matrices D_1 and D_2 that share the same scatter matrix $D_1^T D_1 = D_2^T D_2$ but are otherwise different. We aim to show that the columns of one are reflections of the other and vice versa. Show that a partially shared (full) singular value decomposition can be found for D_1 and D_2 , so that $D_1 = Q_1 \Sigma P^T$ and $D_2 = Q_2 \Sigma P^T$. Use this fact to show that $D_2 = Q_{12} D_1$ for some orthogonal matrix Q_{12} .

Since the scatter matrix is the same, we will have the same right eigenvectors and corresponding eigenvalues in the two cases. We can use these eigenvectors and eigenvalues in order to construct Σ and P in the two cases. The values of Q_1 and Q_2 will be obtained as $Q_1 \Sigma = D_1 P$ and $Q_2 \Sigma = D_2 P$. Note that it is possible that some of the columns of Q corresponding to tied and zero singular values might not be unique, but it does not affect the result as long as the same eigenvectors are chosen in the two cases. Because of the nature of the SVD, it is evident that $D_2 = Q_2 \Sigma P^T = Q_2 Q_1^T (Q_1 \Sigma P^T) = Q_2 Q_1^T D_1$. Therefore, Q_{12} is chosen as $Q_2 Q_1^T$.

20. Let $A = \bar{a}\bar{b}^T$ be a rank-1 matrix for vectors $\bar{a}, \bar{b} \in \mathcal{R}^n$. Find the nonzero eigenvectors, eigenvalues, singular vectors, and singular values of A .

The left and right singular vectors are the unit normalized vectors \bar{a} and \bar{b} , respectively. The only nonzero singular value is the product of the lengths of \bar{a} and \bar{b} . Interestingly, the right eigenvector is the left singular vector and the left eigenvector is the right singular vector. For example, when \bar{a} is used is the right eigenvector, we obtain $A\bar{a} = (\bar{a} \cdot \bar{b})\bar{a}$. Therefore, the only nonzero eigenvalue is $\bar{a} \cdot \bar{b}$.

21. What are the singular values of (i) a $d \times d$ Givens rotation matrix, (ii) a $d \times d$ Householder reflection matrix, (iii) a $d \times d$ projection matrix of rank r , (iv) a 2×2 shear matrix with 1s along the diagonal, and a value of 2 in the upper right corner.

(i) and (ii) All singular values are 1 for orthogonal matrices, (iii) There are r singular values of 1 and the remaining $D - r$ singular values are 0, (iv) One can compute $A^T A$ as a matrix with 1 and 5 on the diagonal entries, and values of 2 on each of the off-diagonal entries. The characteristic polynomial is $\lambda^2 - 6\lambda + 1$, which yields roots of $3 \pm 2\sqrt{2}$. These are the squares of the singular values. The actual singular values are therefore $\sqrt{3 \pm 2\sqrt{2}}$.

22. Consider an $n \times d$ matrix A with linearly independent columns and nonzero singular values $\sigma_1 \dots \sigma_d$. Find the nonzero singular values of $A^T(AA^T)^5$, $A^T(AA^T)^5A$, $A(A^TA)^{-2}A^T$, and $A(A^TA)^{-1}A^T$. Do you recognize the last of these matrices? Which of these matrices have economy SVDs with (additional) zero singular values?

The results can be shown by substituting the economy SVD $Q\Sigma P^T$ in each case. Note that Q is an $n \times d$ matrix in the economy SVD. We obtain the following:

$$A^T(AA^T)^5 = P\Sigma^{11}Q^T$$

Singular values are 11th powers of original

No additional zero singular values

$$A^T(AA^T)^5A = P\Sigma^{12}P^T$$

Singular values are 12th powers of original

No additional zero singular values

$$A(A^TA)^{-2}A^T = Q\Sigma^{-2}Q^T$$

Singular values are negative second powers of the original

However, there are additional zero singular values on the diagonal as this is a larger $n \times n$ matrix. So we will have to augment the $n \times d$ matrix Q with additional columns to create the economy SVD. These columns are any set of orthogonal columns to the ones already included.

$$A(A^TA)^{-1}A^T = QQ^T$$

Nonzero singular values are 1s.

However, there are additional zero singular values on the diagonal as this is a larger $n \times n$ matrix. So we will have to augment the $n \times d$ matrix Q with additional columns to create the economy SVD.

This is the projection matrix.

23. Suppose that you have the $n \times 3$ scatterplot matrix D of an ellipsoid in 3-dimensions, whose three axes have lengths 3, 2, and 1, respectively. The axes directions of this ellipsoid are $[1, 1, 0]$, $[1, -1, 0]$, and $[0, 0, 1]$. You multiply the scatter plot matrix D with a 3×3 transformation matrix A to obtain the scatter plot $D' = DA$ of a new ellipsoid, in which the axes $[1, 1, 1]$, $[1, -2, 1]$, and $[1, 0, -1]$ have lengths 12, 6, and 5, respectively. Write the singular value decompositions of two possible matrices that can perform the transformation. You should be able to write down the SVDs with very little numerical calculation.

There are an infinite number of ways to perform the transformation, although we will perform the axis-to-axis transform. The 3×3 matrix Q is constructed using the columns $[1, 1, 0]$, $[1, -1, 0]$, and $[0, 0, 1]$ in that specific order. Subsequently, we map the axes to one another, which can be done in 6 possible ways. Correspondingly the entries of Σ are $a/3$, $b/2$, and c , where a , b , and c are 12, 6, and 5 in any order. Let us say that we select the order 12, 5, and 6, so that the entries of Σ are 4, 2.5, and 6. In such a case, the matrix P has columns in the order $[1, 1, 1]$, $[1, 0, -1]$, and $[1, -2, 1]$.

24. **Regularization impact:** Consider the regularized least-squares regression problem of minimizing $\|A\bar{x} - \bar{b}\|^2 + \lambda\|\bar{x}\|^2$ for d -dimensional optimization vector \bar{x} , n -dimensional vector \bar{b} , nonnegative scalar λ , and $n \times d$ matrix A . There are several ways of showing that the norm of the optimum solution $\bar{x} = \bar{x}^*$ is non-increasing with increasing λ (and

this is also intuitively clear from the nature of the optimization formulation). Use SVD to show that the optimum solution $\bar{x}^* = (A^T A + \lambda I_d)^{-1} A^T \bar{b}$ has non-increasing norm with increasing λ .

Substitute $A = Q\Sigma P^T$ using the compact SVD, and show that the expression simplifies to the norm of $P(\Sigma^2 + \lambda I)^{-1} \Sigma Q^T \bar{b}$. We can ignore the leading P because the rotation does not affect the norm and substitute $\bar{y} = Q^T \bar{b}$. Therefore, the expression simplifies to the norm of $(\Sigma^2 + \lambda I)^{-1} \Sigma \bar{y}$. We are effectively scaling each component of \bar{y} by a fraction that reduces with increasing λ before computing its norm. The overall expression is the same as that in the statement of Exercise 25. Therefore, the expression has non-increasing norm with increasing λ .

25. The function $f(\lambda)$ arises commonly in spherically constrained least-squares regression:

$$f(\lambda) = \bar{b}^T A (A^T A + \lambda I)^{-2} A^T \bar{b}$$

Here, A is an $n \times d$ matrix of rank- r , \bar{b} is an n -dimensional column vector, and $\lambda > 0$ is an optimization parameter. Furthermore, $A = Q\Sigma P^T$ is the reduced SVD of A with $n \times r$ matrix Q , $d \times r$ matrix P , and $r \times r$ diagonal matrix Σ . The diagonal elements of Σ are $\sigma_{11} \dots \sigma_{rr}$. Show that $f(\lambda)$ can be written in scalar form as follows:

$$f(\lambda) = \sum_{i=1}^r \left(\frac{\sigma_{ii} c_i}{\sigma_{ii}^2 + \lambda} \right)^2$$

Here, c_i is the i th component of $Q^T \bar{b}$.

This exercise is very similar to the previous one, because we are essentially computing the norm of the solution to least-squares regression.

26. Pseudoinverse properties: Show using SVD that $AA^+A = A$ and $A^+AA^+ = A^+$. Also show using SVD that AA^+ is a symmetric and idempotent matrix (which is an alternative definition of a projection matrix).

In each case, we substitute $A = Q\Sigma^{-1}P^T$ using the compact SVD of A and simplify the underlying expressions using $P^T P = Q^T Q = I$. However, note that QQ^T and PP^T are not equal to I . For example, it can be shown that $AA^+A = Q\Sigma P^T = A$. Similarly, we have $A^+AA^+ = Q\Sigma^{-1}P^T = A^+$. We can also compute $AA^+ = QQ^T$, which is both symmetric and idempotent. In fact, QQ^T is the projection matrix onto the column space of A .

27. Compute the compact SVD of the matrix A :

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 2 & 0 \end{bmatrix}$$

The SVD of the matrix is as follows:

$$Q = \begin{bmatrix} -0.9347217 & -0.35538056 \\ -0.35538056 & 0.9347217 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 3.9396 & 0 \\ 0 & 1.865 \end{bmatrix}$$

$$P^T = \begin{bmatrix} -0.56472711 & -0.41767294 & -0.71178129 \\ 0.12006923 & 0.81171587 & -0.5715774 \end{bmatrix}$$

Note that this decomposition can be computed by first computing Q , which is smaller than the matrix P . The matrix Q can be computed as the eigenvectors of DD^T . This yields the matrix Q and Σ . Then, the matrix $D^T Q$ yields $P\Sigma$.

The Moore-Penrose pseudo-inverse is given by $Q\Sigma^+P^T$, which is as follows:

$$A^+ = \begin{bmatrix} 1/9 & 1/9 \\ -1/18 & 4/9 \\ 5/18 & -2/9 \end{bmatrix}$$

28. Generalized singular value decomposition: *The generalized singular value decomposition of an $n \times d$ matrix D is given by $D = Q\Sigma P^T$, where $Q^T S_1 Q = I$ and $P^T S_2 P = I$. Here, S_1 and S_2 are (given) $n \times n$ and $d \times d$ positive definite matrices, and therefore the singular vectors are orthogonal from the perspective of the generalized definition of inner products. Show how to reduce generalized singular value decomposition to a singular value decomposition on a modified version of D .*

Define $D' = \sqrt{S_1} D \sqrt{S_2}$. As discussed in Chapter 3, a symmetric square-root matrix can always be extracted from positive definite matrices using the same eigenvectors and the square-roots of the eigenvalues to construct the square-root matrix. Then, the SVD of D' is as follows:

$$D' = U\Sigma V^T$$

Then, we set Q and P as follows:

$$Q = (S_1)^{-1/2} U, \quad P = (S_2)^{-1/2} V$$

Therefore, we have $D' = U\Sigma V^T = S_1^{1/2} Q \Sigma P^T S_2^{1/2}$. Therefore, we have $Q \Sigma P^T = S_1^{-1/2} D' S_2^{-1/2} = D$.

Chapter 8

Matrix Factorization

1. **Biased matrix factorization:** Consider a situation in which the incomplete $n \times d$ matrix D is approximately factorized into an $n \times k$ matrix U and a $d \times k$ matrix V for prediction as follows:

$$D \approx UV^T$$

Now suppose that you add the constraint that all entries of the penultimate column of U and the final column of V are fixed to 1 during learning. Discuss the similarity of this model with the addition of bias to classification models. How is gradient-descent modified?

The values of the penultimate columns of U and V represent the biases for the corresponding factors. For example, in a recommendation application, some users are more likely to give high ratings and some items are likely to get rated highly. The bias values correspond to these differences. The updates are the same as traditional gradient descent, except that we reset the corresponding columns of U and V after each update to 1.

2. In the scenario of Exercise 1, will the Frobenius norm on observed ratings be better optimized with or without constraints on the final columns of U and V ? Why might it be desirable to add such a constraint during the estimation of missing entries?

The error on the observed entries of the matrix will be higher by adding constraints. However, the approach will generalize better to unseen entries.

3. Suppose that you have a symmetric $n \times n$ matrix D of similarities, which has missing entries. You decide to recover the missing entries by using the symmetric factorization $D \approx UU^T$. Here, U is an $n \times k$ matrix, and k is the rank of the factorization.

- (a) Write the objective function for the optimization model using the Frobenius norm and L_2 -regularization.
- (b) Derive the gradient-descent steps in terms of matrix-centric updates.
- (c) Discuss the conditions under which an exact factorization will not exist, irrespective of how large a value of k is used for the factorization.

The objective function is $J = \frac{1}{2} \|D - UU^T\|^2 + \frac{\lambda}{2} \|U\|^2$.

The update is $U \leftarrow U(1 - \alpha\lambda) + 2\alpha EU$

The main issue is whether D is positive semi-definite. This is because any matrix of the form UU^T must be positive semidefinite. Note that any symmetric matrix can be made semi-definite by adding a sufficient amount to the diagonal entries.

4. *Derive the gradient-descent updates for L_1 -loss matrix factorization in which the objective function is $J = \|D - UV^T\|_1$.*

One can show using similar steps as L_2 -loss that the following partial derivatives are appropriate:

$$\begin{aligned}\frac{\partial J}{\partial U} &= -\text{sign}((D - UV^T))V = -\text{sign}(E)V \\ \frac{\partial J}{\partial V} &= -\text{sign}((D - UV^T)^T)U = -\text{sign}(E^T)U\end{aligned}$$

Here, the sign function is applied in elementwise fashion to the matrix to create a matrix of the same size. The corresponding gradient-descent updates are as follows:

$$\begin{aligned}U &\Leftarrow U - \alpha \frac{\partial J}{\partial U} = U + \alpha \text{sign}(E)V \\ V &\Leftarrow V - \alpha \frac{\partial J}{\partial V} = V + \alpha \text{sign}(E^T)U\end{aligned}$$

5. *Derive the gradient-descent updates for L_2 -loss matrix factorization in which L_1 -regularization is used on the factors.*

Using the same notations as the previous exercise, the updates are as follows:

$$\begin{aligned}U &\Leftarrow U - \alpha \frac{\partial J}{\partial U} = U - \alpha \lambda \text{sign}(U) + \alpha (E)V \\ V &\Leftarrow V - \alpha \frac{\partial J}{\partial V} = V - \alpha \lambda \text{sign}(V) + \alpha (E^T)U\end{aligned}$$

6. *In SVD, it is easy to compute the representation of out-of-sample matrices because of the orthonormality of the basis $d \times k$ matrix V . If the SVD factorization of the $n \times d$ matrix D is $D \approx UV^T$, then one can compute the representation of an out-of-sample $m \times d$ matrix D_o as $D_o V$. Show how you can efficiently compute a similar out-of-sample representation of D_o , when you are given a non-orthonormal factorization $D = UV^T$. Assume that m and k are much smaller than n and d . [Hint: This trick has implicitly been used as a subroutine of one of the methods discussed in this chapter for matrix factorization.]*

The out-of-sample embedding is given by $D_o V (V^T V)^{-1}$. Note that this result can be derived by reducing the problem to a sequence of similar linear regression problems on rows of D_o and corresponding rows of U .

7. *Show that the k -means formulation in Chapter 3 is identical to the formulation of this chapter.*

The matrix U contains the assignment variables of the points to each cluster, which are the variables y_{ij} in Chapter 3. Therefore, the variables y_{ij} are equivalent to the variables u_{ij} in this chapter. The j th column of matrix V contains the centroid of the cluster, which is the variable vector \bar{z}_j in Chapter 3. Note that the columns of U can be mutually orthogonal only when two columns in the same row never contain a

1. In other words, each row contains a single 1 in U . This is the same constraint as in Chapter 3. Furthermore, the objective function of aggregate mean-squared error is the same in the two cases.

- 8. Orthogonal Nonnegative Matrix Factorization:** Consider a nonnegative $n \times d$ data matrix D in which we try to approximately factorize D as UV^T with the Frobenius norm as the objective function. Suppose you add nonnegativity constraints on U and V along with the constraint $U^T U = I$. How many entries in each row of U will be nonzero? Discuss how you can extract a clustering from this factorization. Show that this approach is closely related to the k -means optimization formulation.

At most one entry in each row is nonzero. However unlike the previous exercise, that value can be any real value. Therefore, when one is computing the squared error, one is computing the distance of the point to a scaled version of the mean. In a sense, this approach is a relaxed version of the k -means algorithm. Note that it is not an integer program, and therefore it is easier to solve than the previous exercise.

- 9.** Suppose that you use GloVe on a quantity matrix $Q = [q_{ij}]$ in which each count q_{ij} is either 0 or 10000. A sizeable number of counts are 0s. Show that GloVe can discover a trivial factorization with zero error in which each word has the same embedded representation.

The reason for this is discussed in the text of the book. The basic point is that negative sampling is important. GloVe does not perform negative sampling, and it is susceptible in cases where the nonzero frequencies are very similar. In such cases, it is unable to distinguish between nonzero and zero frequencies via negative sampling.

- 10.** Derive the gradient update equations for using factorization machines in binary classification with logistic loss and hinge loss.

The gradient descent updates can be derived by using the same approach as the squared loss except that the derivative of the logistic function or the hinge function needs to be multiplied within the chain rule, instead of the difference between observed and predicted value.

- 11.** Suppose you want to perform the rank- k factorization $D \approx UV^T$ of the $n \times d$ matrix D using gradient descent. Propose an initialization method for U and V using QR decomposition of k randomly chosen columns of D .

Let D_k be a $n \times k$ matrix obtained by sampling k columns of D . Without loss of generality assume that these are the first k columns of $D = [D_k D_{n-k}]$ (since we can always shuffle columns of D after the fact along with the corresponding rows of V). Perform the QR decomposition of $D_k = Q_k R_k$. Set $U = Q_k$. Create the matrix V^T as $[R_k, Q_k^T D_{n-k}]$.

- 12.** Suppose that you have a sparse non-negative matrix D of size $n \times d$. What can you say about the dot product of any pair of columns as a consequence of sparsity? Use this fact along with the intuition derived from the previous exercise to initialize U using k randomly sampled columns of D for non-negative matrix factorization. In this case, the initialized matrices U and V need to be non-negative.

The columns of D will be roughly orthogonal. Therefore, U can be initialized to k randomly chosen columns of D , and then the columns can be normalized to one unit. Subsequently, the matrix V can be initialized as $D^T U$.

- 13. Nonlinear matrix factorization of positive matrices:** Consider a nonlinear model for matrix factorization of positive matrices $D = [x_{ij}]$, where $D = F(UV^T)$, and $F(x) = x^2$ is applied in element-wise fashion. The vectors \bar{u}_i and \bar{v}_j represent the i th and j th rows of U and V , respectively. The loss function is $\|D - F(UV^T)\|_F^2$. Show that the gradient descent steps are as follows:

$$\bar{u}_i \leftarrow \bar{u}_i + \alpha \sum_j (\bar{u}_i \cdot \bar{v}_j)(x_{ij} - F(\bar{u}_i \cdot \bar{v}_j))\bar{v}_j, \quad \bar{v}_j \leftarrow \bar{v}_j + \alpha \sum_i (\bar{u}_i \cdot \bar{v}_j)(x_{ij} - F(\bar{u}_i \cdot \bar{v}_j))\bar{u}_i$$

One can solve this problem with the use of an application of the chain rule. Here, the key point is that the derivative of $F(x^2)$ is simply $2x$, which amounts to $\bar{u}_i \cdot \bar{v}_j$ when the function is applied to entry x_{ij} . The constant factor of 2 can be absorbed in the learning rate. Therefore, the gradient descent update is identical to the update in unconstrained matrix factorization (discussed in chapter) with an additional factor of $\bar{u}_i \cdot \bar{v}_j$ within the updates. This is exactly what is shown in the update equations.

- 14. Out-of-sample factor learning:** Suppose that you learn the optimal matrix factorization $D \approx UV^T$ of $n \times d$ matrix D , where U, V are $n \times k$ and $d \times k$ matrices, respectively. Now you are given a new out-of-sample $t \times d$ data matrix D_o with rows collected using the same methodology as the rows of D (and with the same d attributes). You are asked to quickly factorize this out-of-sample data matrix into $D_o \approx U_o V^T$ with the objective of minimizing $\|D_o - U_o V^T\|_F^2$, where V is fixed to the matrix learned from the earlier in-sample factorization. Show that the problem can be decomposed into t linear regression problems, and the optimal solution U_o is given by:

$$U_o^T = V^+ D_o^T$$

Here, V^+ is the pseudoinverse of V . Show that the rank- k approximation of $D_o \approx U_o V^T$ is given by $D_o P_v$, where $P_v = V(V^T V)^{-1} V^T$ is the $d \times d$ projection matrix induced by V . Propose a fast solution approach using QR decomposition of V and back-substitution with a triangular equation system. How does this problem relate to the alternating minimization approach?

The Frobenius norm can be decomposed into the squared norms for each of the rows of D_o . This is a linearly separable objective function in which the parameters for the t th row of U_o are \bar{w}_t (which is a k -dimensional vector). The linear regression factors \bar{w}_t for each of the t rows of U_o are learned using this approach. Therefore, if the t th row of D_o is \bar{y}_t , the problem boils down to the linear regression problem $V \bar{w}_t^T = \bar{y}_t^T$. The solution is based on the pseudo-inverse $\bar{w}_t^T = V^+ \bar{y}_t^T$. The same condition holds for each row of D_o . Therefore, we obtain the following:

$$U_o^T = V^+ D_o^T$$

In alternating least-squares, one alternately learns U and V with this approach, whereas we only learn one of the factor matrices here. One can use QR decomposition and backsubstitution in a manner that is similar to what is described in the text on linear regression.

- 15. Out-of-sample factor learning:** Consider the same scenario as Exercise 14, where you are trying to learn the out-of-sample factor matrix U_o for in-sample data matrix

$D \approx UV^T$ and out-of-sample data matrix D_o . The factor matrix V is fixed from in-sample learning. Closed-form solutions, such as the one in Exercise 14, are rare in most matrix factorization settings. Discuss how the gradient-descent updates discussed in this chapter can be modified so that U_o can be learned directly. Specifically discuss the case of (i) unconstrained matrix factorization, (ii) nonnegative matrix factorization, and (iii) logistic matrix factorization.

In each case, we perform only the updates for U repeatedly while keeping V fixed.

16. Suppose that you have a data set in which users have rated small subsets of items with a numerical value. Furthermore, users also specify directed trust and distrust links to each other in order to show their confidence in each other's feedback.

- (a) Show how you can use shared matrix factorization for estimating the rating of a user on an item that they have not already rated.
- (b) Show how you can use factorization machines to achieve similar goals as (a).

(a) We have two sets of user factors, U and Z , and one set of item vectors V . The ratings matrix is given by $R \approx UV^T$ and $R \approx ZV^T$. Furthermore, the trust matrix is given by $T \approx ZU^T$. Then, we can set up an optimization problem in order to find U , V , and Z .

(b) In this case, we set up separate attributes for source users, destination users and items. Each row contains exactly three 1s. The target variable is the rating.

17. Propose an algorithm for finding outlier entries in a matrix with the use of matrix factorization.

The absolute values of residuals of the matrix factorization provide the outlier scores. Entries with large absolute values of the residuals are reported as the outlier entries.

18. Suppose that you are given the linkage of a large Website with n pages, in which each page contains a bag of words drawn from a lexicon of size d . Furthermore, you are given information on how m users have rated each page on a scale of 1 to 5. The ratings data is incomplete. Propose a model to create an embedding for each Webpage by combining all three pieces of information.

Let A be the $n \times n$ matrix for the Web graph, D be the $n \times d$ document-term matrix for the documents associated with pages, and R be the $n \times m$ matrix for the ratings that the users have associated with the Web pages. Then, the following shared matrix factorization model can be proposed:

$$\begin{aligned} A &\approx UV^T \\ D &\approx UW^T \\ R &\approx UZ^T \end{aligned}$$

Each matrix has rank k and the sizes of the matrices are chosen to be consistent. The matrix U is a shared factor, and its rows will provide an embedding of the various Web pages. This embedding includes information from all the modalities. One can set up a model that minimizes the sum of the Frobenius norms of the residuals of the different matrix factorizations. The different terms can be weighted differently, if different modalities are given different importance. Furthermore, only the observed entries of R are used in the error portion of the model.

- 19. True or false:** A zero error non-negative matrix factorization (NMF) UV^T of an $n \times d$ non-negative matrix D always exists, where U is an $n \times k$ matrix and V is a $d \times k$ matrix, as long as k is chosen large enough. At what value of k can you get an exact NMF of the following matrix?

$$D = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

The answer is “true.” We can decompose any matrix multiplication UV^T into the sum of outer products. As long as D can be decomposed into the sum of non-negative rank-1 matrices, one can obtain a factorization. For example, we can set $U = D$ and $V = I_d$, or we can set $U = I_n$ and $V = D^T$. Therefore, at $k = \min\{n, d\}$, the matrix can be reconstructed. Therefore, we can reconstruct each row (or column) one by one and add up the rank-1 matrices. Although this decompositions might seem trivial and one of the factors looks like a rectangular version of the identity matrix, this is not the only decomposition. As shown in the next exercise, alternative non-trivial decompositions can be reconstructed from these basic decompositions.

The matrix D can be decomposed into rank-1 matrices as follows:

$$D = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0.5 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.5 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 0.5 & 0 \\ 1 & 2 \end{bmatrix}$$

- 20. True or false:** Suppose you have the exact non-negative factorization (NMF) UV^T of a matrix D , so that each column of V is constrained to sum to 1. Subject to this normalization rule, the NMF of D is unique.

The answer is “false.” Given a factorization UV^T , we can pick any $k \times k$ permutation matrix P to create $UV^T = (UP)(VP)^T$.

- 21.** Discuss why the following algorithm will work in computing the matrix factorization $D_{n \times d} \approx UV^T$ after initializing $U_{n \times k}$ and $V_{d \times k}$ randomly:

repeat; $U \leftarrow DV^+$; $V \leftarrow D^T U^+$; **until convergence;**

This is simply the alternating least-squares algorithm (block coordinate descent), which is implemented with the help of the Moore-Penrose pseudoinverse.

- 22.** Derive the gradient-descent updates of unconstrained matrix factorization with L_1 -regularization. You may assume that the regularization parameter is $\lambda > 0$.

$$U \leftarrow U - \alpha \frac{\partial J}{\partial U} = U + \alpha EV - \alpha \lambda \text{sign}(U)$$

$$V \leftarrow V - \alpha \frac{\partial J}{\partial V} = V + \alpha E^T U - \alpha \lambda \text{sign}(V)$$

Here, the sign function is applied in entry-wise fashion. Note that the only part of the gradient-descent update that is different from the case of L_2 -regularization is in terms of how the shrinkage is done.

- 23. Alternating nonnegative least-squares:** *Propose an algorithm for nonnegative matrix factorization using the alternating least-squares method.*

This is similar to unconstrained matrix factorization, except that nonnegative box regression needs to be applied during the alternating iterations. Nonnegative box regression is discussed in Chapter 6.

- 24. Bounded matrix factorization:** *In bounded matrix factorization, the entries of U and V in the factorization $D \approx UV^T$ are bounded above and below by specific values. Propose a computational algorithm for bounded matrix factorization using (i) gradient descent, and (ii) alternating least-squares.*

(i) Gradient descent is the same as unconstrained matrix factorization, except that the entries of U and V need to be reset to their bounds after each iteration. (ii) Alternating least-squares is the same as unconstrained matrix factorization, except that box regression needs to be applied in each iteration instead of unconstrained regression. Box regression is discussed in Chapter 6.

- 25.** *Suppose that you have a very large and dense matrix D of low rank that you cannot hold in memory, and you want to factorize it as $D \approx UV^T$. Propose a method for factorization that uses only sparse matrix multiplication.*

In recommender systems, one can use the observed entries in order to perform the following updates with sparse matrix multiplications (see text book to check how the error matrix E is constructed):

$$\begin{aligned} U &\Leftarrow U - \alpha \frac{\partial J}{\partial U} = U(1 - \alpha\lambda) + \alpha EV \\ V &\Leftarrow V - \alpha \frac{\partial J}{\partial V} = V(1 - \alpha\lambda) + \alpha E^T U \end{aligned}$$

The main point is that E contains non-zero entries for only observed data in the recommender matrix. In this case, we can repeatedly sample entries in order to create the “observed” entries. Unlike recommender systems, the sampled entries can be different in each iteration.

- 26. Temporal matrix factorization:** *Consider a sequence of $n \times d$ matrices $D_1 \dots D_t$ that are slowly evolving over t time stamps. Show how one can create an optimization model to infer a single $n \times k$ static factor matrix that does not change over time, and multiple $d \times k$ dynamic factor matrices, each of which is time-specific. Derive the gradient descent steps to find the factor matrices.*

The matrix factorizations are as follows:

$$D_i = UV_i^T \quad \forall i$$

Therefore, the objective function is $\sum_i \|D_i - UV_i^T\|^2$. We set the error matrices to $E_i = D_i - UV_i^T$ for each i . Then, the updates are as follows:

$$\begin{aligned} U &\Leftarrow U + \alpha \sum_i E_i V \\ V_i &\Leftarrow V_i + \alpha E_i^T U \quad \forall i \end{aligned}$$

In the event that regularization is used, the updates are as follows:

$$\begin{aligned}U &\Leftarrow U(1 - \alpha\lambda) + \alpha \sum_i E_i V \\V_i &\Leftarrow V_i(1 - \alpha\lambda) + \alpha E_i^T U \quad \forall i\end{aligned}$$

Chapter 9

The Linear Algebra of Similarity

1. Suppose that you are given a 10×10 binary matrix of similarities between objects. The similarities between all pairs of objects for the first four objects is 1, and also between all pairs of objects for the next six objects is 1. All other similarities are 0. Derive an embedding of each object.

One can create a 2-dimensional embedding in which the first four objects have the embedding $(1, 0)$ and the last six objects have the embedding $(0, 1)$. This can also be obtained using eigendecomposition of the underlying similarity matrix and taking the top-2 eigenvectors.

2. Suppose that you have two non-disjoint sets of objects A and B . The set $A \cap B$ is a modestly large sample of objects. You are given all similarities between pairs of objects, one drawn from each of the two sets. Discuss how you can efficiently approximate the entire similarity matrix over the entire set $A \cup B$. It is known that the similarity matrix is symmetric.

The main issue is that we know the similarities between the pairs A and B , but not similarities between two objects from A or between two objects from B . Let U be the $n_A \times k$ matrix containing embedding of A , and V be the $n_B \times k$ matrix containing the embedding of B . Then UV^T is the given matrix of similarities S . Then, we can learn U and V by using matrix factorization $S \approx UV^T$. This approach is described in Chapter 7 for recommender systems (which compute similarities between users and items). Subsequently, one can use UU^T and VV^T to find the similarities within the sets A and B , respectively.

3. Suppose that S_1 and S_2 are positive semi-definite matrices of ranks k_1 and k_2 , respectively, where $k_2 > k_1$. Show that $S_1 - S_2$ can never be positive semi-definite.

Since, S_2 has larger rank than S_1 , the null space of S_1 has larger dimensionality than the null space of S_2 . Therefore, a vector \bar{x} must exist in the null space of S_1 that is not present in the null space of S_2 . This vector will satisfy $S_2\bar{x} \neq \bar{0}$ and $S_1\bar{x} = \bar{0}$. Therefore, we will have $\bar{x}^T S_1 \bar{x} = 0$. For PSD matrices $S_2\bar{x} \neq \bar{0}$ implies that $\bar{x}^T S_2 \bar{x} > 0$. In other words, we will have $\bar{x}^T (S_1 - S_2) \bar{x} < 0$. This is not possible for PSD matrices.

We want to perform the matrix factorization $S \approx UU^T$ with only a subset of the entries in S . One can learn U by performing stochastic gradient descent on only the specified entries.

4. Suppose you are given a binary matrix of similarities between objects, in which most entries are 0s. Discuss how you can adapt the logistic matrix factorization approach of Chapter 7 to make it more suitable to symmetric matrices.

Note that we can use the parameter sharing trick discussed in Chapter 3 and also directly use the results in Chapter 7, where the matrix factorization UV^T is assumed and gradients are computed separately with respect to U and V . Then, we can add the gradients with respect to U and with respect to V .

5. Suppose that you were given an incomplete matrix of similarities between objects belonging to two sets A and B that are completely disjoint. Discuss how you can find an embedding for each of the objects in the two sets. Are the embeddings of the objects in set A comparable to those in the set B ?

This is simply an asymmetric matrix factorization in which one can use stochastic gradient descent only over specified items. The setting is very similar to that of recommender systems. The embeddings U and V are not directly comparable without additional assumptions or regularizations (e.g., magnitudes of entries in U and V are similar). For example, one can multiply U with 2 and V with 0.5 without changing the result.

6. A centered vector is one whose elements sum to 0. Show that for any valid (squared) distance matrix $\Delta = [\delta_{ij}^2]$ defined on a Euclidean space, the following must be true for any centered d -dimensional vector \bar{y} :

$$\bar{y}^T \Delta \bar{y} \leq 0$$

- (a) Suppose that you are given a symmetric matrix Δ in which all entries along the diagonal are 0s, and it satisfies the condition $\bar{y}^T \Delta \bar{y} \leq 0$ for any centered vector \bar{y} . Show that all entries of Δ must be nonnegative by using an appropriate choice of vector \bar{y} .
- (b) Discuss why a distance matrix Δ of (squared) Euclidean distances is always indefinite, unless it is a trivial matrix of 0s.

Note that the similarity matrix S can be expressed as $-(I - M/n)\Delta(I - M/n)/2$, where M is a matrix of 1s. For any centered vector \bar{y} one can show that $\bar{y}^T S \bar{y}$ is equal to $-\bar{y}^T \Delta \bar{y}/2$. Since we know that $\bar{y}^T S \bar{y}$ is nonnegative, it follows that $\bar{y}^T \Delta \bar{y}$ is non-positive.

- (a) This can be easily shown by picking a vector in which the i th and j th elements are 1 and -1, respectively.
 - (b) The eigenvalues must sum to zero, which is the trace of the distance matrix. Unless all eigenvalues are zero (trivial matrix), the matrix is indefinite.
7. You have an $n \times n$ (dot-product) similarity matrix between training points and a $t \times n$ similarity matrix S_t between test and training points. The n -dimensional column vector of class variables is \bar{y} . Furthermore, the true $n \times d$ data matrix is D (i.e., $S = DD^T$),

but you are not shown this matrix. The d -dimensional coefficient vector \bar{W} of linear regression is given by the following:

$$\bar{W} = (D^T D + \lambda I)^{-1} D^T \bar{y}$$

Here, λ is the regularization parameter. Then, show the following results:

- (a) Let \bar{p} be the t -dimensional vector of predictions for test instances. Show the following using the Woodbury identity (variation):

$$\bar{p} = S_t(S + \lambda I)^{-1} \bar{y}$$

- (b) Show the result of (a) using the representer theorem discussed in this chapter.
(c) Take a moment to examine the coefficient vector obtained using the dual approach. What do you observe?

- (a) Using the Woodbury inequality we can get the following vector:

$$\bar{W} = D^T(DD^T + \lambda I)^{-1} \bar{y} = D^T(S + \lambda I)^{-1} \bar{y}$$

If the unknown test matrix is D_t , the prediction for the test instances is given by the vector $D_t \bar{W}$. Using the fact that $D_t D^T = S_t$, we can derive the result.

- (b) By setting the gradient of the objective function to 0, one can obtain the coefficient of the representer parameters as follows:

$$\bar{\beta} = (S + \lambda I)^{-1} \bar{y}$$

Since the predictions are given by $S_t \beta$, we obtain the result.

8. Derive the gradient descent steps for the primal formulation of logistic regression using the similarity matrix S and the representer theorem.
9. A student is given a square and symmetric similarity matrix S that is not positive semi-definite. The student computes the following new matrix:

$$S' = I - S + S^2$$

Is the new similarity matrix S' always positive semi-definite? If it is positive semi-definite, provide a proof. Otherwise, provide a counterexample.

Yes, the new matrix is always PSD. Express the symmetric matrix as $P \Delta P^T$. Then, the matrix S' can be expressed as $P(\Delta^2 - \Delta + I)P^T$. The diagonal matrix only has positive entries on the diagonal. This is because each entry is of the form $x^2 - x + 1 = (x - 0.5)^2 + 0.75$. Therefore, the matrix is not only PSD, but it is positive definite.

10. A student used three different ways to estimate $n \times n$ similarity matrices S_1 , S_2 , and S_3 among a set of n objects. These similarity matrices were all positive semi-definite. The student then computed the composite similarity matrix S as follows:

$$S = S_1 \odot S_2 + S_2 \odot S_3 + S_3 \odot S_1$$

Is the composite similarity matrix positive semi-definite?

Each matrix of the for $S_i \odot S_j$ is PSD based on the results discussed in the text. Furthermore, the sum of PSD matrices is PSD.

11. Suppose $S(\bar{X}_1, \bar{X}_2) = S(\bar{X}_2, \bar{X}_1)$ is a symmetric similarity function between vectors \bar{X}_1 and \bar{X}_2 , which is not necessarily a valid kernel. Then, is the similarity function $K(\bar{X}_1, \bar{X}_2) = S(\bar{X}_1, \bar{X}_2)^2$ a valid kernel?

Not necessarily. Consider a matrix in which the diagonal elements are 0s. Such a matrix will always have negative eigenvalues unless it is the zero matrix, because the trace sums to 0. On applying the above function, the diagonal entries continue to be 0s.

12. Suppose that S is a positive semi-definite kernel, and a sub-linear element-wise function $f(\cdot)$ is applied to each element of S to create the new matrix $f(S)$. In each case, either show that $f(S)$ is positive semi-definite or provide a counter-example: (i) $f(x)$ is the natural logarithmic function, and S originally contains positive entries, and (ii) $f(x)$ is the non-negative square-root function, and S originally contains nonnegative entries.

Neither of the above statements are true. Counter-examples of the matrix S are as follows:

- (i) Consider the following matrix S :

$$S = \begin{bmatrix} 1 & 1/e \\ 1/e & 1 \end{bmatrix}$$

This matrix is PSD because the sum of eigenvalues (trace) is positive and so is the product of eigenvalues (determinant). Therefore, both eigenvalues are positive. In this case, the matrix $f(S)$ is as follows:

$$f(S) = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

Here, the sum of eigenvalues of 0, and the product of eigenvalue is -1 . Therefore, the two eigenvalues are 1 and -1 , which is indefinite.

- (ii) This case is much harder. In this case, consider the matrix S as follows:

$$S = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

This matrix has eigenvalues of 1, $2 + \sqrt{2}$ and $2 - \sqrt{2}$, all of which are nonnegative. Therefore, the matrix is PSD. Now consider what happens after applying the element-wise function:

$$f(S) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & \sqrt{3} \end{bmatrix}$$

This matrix has a determinant (product of eigenvalues) of $\sqrt{3} - 2$, which is negative. Therefore, the matrix cannot be PSD.

13. **Symmetric nonnegative factorization:** Consider a symmetric and nonnegative $n \times n$ matrix S that is factorized as $S \approx UU^T$, where U is an $n \times k$ matrix for some $k < n$. The errors on the diagonal entries are ignored with the use of the objective

function $\|W \odot (S - UU^T)\|^2$. Here, W is an $n \times n$ binary weight matrix that is set to 1 for all entries other than the diagonal ones. Suppose that you additionally constrain all entries in U to be nonnegative. Derive a projected gradient-descent update for this box-constrained optimization problem. Discuss why the factor matrix U is more interpretable in the nonnegative case.

The updates are the same as symmetric matrix factorization, except that one needs to set each negative entry to 0 after each update.

14. Show that at least one symmetric factorization $S = UU^T$ exists of a positive semi-definite matrix S , so that U is symmetric as well.

Diagonalize $S = Q\Sigma^2Q^T$ and set $U = Q\Sigma Q^T$. This matrix is also referred to as the square-root matrix because $UU^T = U^2 = S$.

15. Express the loss function of the regularized L_2 -loss SVM (cf. Chapter 5) using the representer theorem in terms of a similarity matrix. Here, we will convert the regularized Newton update of Chapter 5 to a representer update. The Newton update of Chapter 5 (using the same notations as the chapter) is as follows:

$$\bar{W} \leftarrow (D_w^T D_w + \lambda I_d)^{-1} D_w^T \bar{y}$$

Here, the $n \times d$ matrix $D_w = \Delta_w D$ is a partial copy of D , except that it has zero rows for margin-satisfying rows of D at current values of the weight vector. Δ_w is a binary diagonal matrix containing values of 1 only for margin-violating rows. Use the Woodbury identity to show that this update is equivalent to the following:

$$\bar{\beta} \leftarrow \Delta_w (S_w + \lambda I_n)^{-1} \bar{y}$$

Here, $S_w = D_w D_w^T$ is a similarity matrix. Note that the update is similar to the closed form of linear regression in the previous exercise except that it sets dependent/independent variables of margin-satisfying points to 0 (and it needs to be iteratively repeated). Explain this similarity.

The loss function is as follows:

$$J = \frac{1}{2} \|\Delta_w (S\bar{\beta} - \bar{y})\|^2 + \frac{\lambda}{2} \bar{\beta}^T S \bar{\beta} \quad (9.1)$$

The value of $y_i(\bar{W} \cdot \bar{X}_i = y_i(\sum_j \beta_j s_{ij})$ needs to be less than 1 for the i th point to be margin violating. The update for \bar{W} can be converted using the Woodbury identity as follows:

$$\bar{W} \leftarrow D_w^T (D_w D_w^T + \lambda I_n)^{-1} \bar{y} = D_w^T (S_w + \lambda I_n)^{-1} \bar{y} = D^T \Delta_w (S_w + \lambda I_n)^{-1} \bar{y}$$

Note that the update of \bar{W} can be expressed by decomposing it in the form $\sum_i \beta_i \bar{X}_i^T$. The value of β_i is 0 if the point is margin satisfying. Otherwise, it is equal to $\sum_j y_j a_{ij}$, where a_{ij} is the (i, j) th entry of the symmetric inverse $A = (S_w + \lambda I_n)^{-1}$ of the symmetric matrix $(S_w + \lambda I_n)$. One can also implicitly make this update by setting β_i to $\sum_j y_j a_{ij}$ for margin-violating points and 0 otherwise. This update is equivalent to the following:

$$\bar{\beta} \leftarrow \Delta_w (S_w + \lambda I_n)^{-1} \bar{y}$$

The similarity between least-squares regression and the L_2 -SVM is discussed in Chapter 4, where it is shown that well-separated points are ignored in the loss function, and the loss function is otherwise identical.

Chapter 10

The Linear Algebra of Graphs

1. Consider the $n \times n$ adjacency matrices A_1 and A_2 of two graphs. Suppose that the graphs are known to be isomorphic. Two graphs are said to be isomorphic, if one graph can be obtained from the other by reordering its vertices. Show that isomorphic graphs have the same eigenvalues. [Hint: What is the nature of the relationship between their adjacency matrices in algebraic form? You may introduce any new matrices as needed.]

Two graphs are isomorphic, if one can express $A_2 = PA_1P^T$, where P is a permutation matrix. Since these matrices are similar, they have the same eigenvalues.

2. Suppose that you were given the eigenvectors and eigenvalues of the stochastic transition matrix P of an undirected graph. Discuss how you can quickly compute P^∞ using these eigenvectors and eigenvalues.

An undirected graph has real eigenvectors and eigenvalues, and is diagonalizable with maximum eigenvalue of 1. Therefore, it can be expressed as $P = V\Delta V^T$. Therefore, P^∞ can be expressed $V\Delta^\infty V^T$. Any eigenvalue that is 1 gets set to 1. Any eigenvalue less than 1 gets set to 0.

3. Let Δ be the weighted degree of matrix of the (undirected) adjacency matrix A , and $\bar{e}_1 \dots \bar{e}_n$ be the n eigenvectors of the stochastic transition matrix $P = \Delta^{-1}A$. Show that any pair of eigenvectors \bar{e}_i and \bar{e}_j are Δ -orthogonal. In other words, any pair of eigenvectors \bar{e}_i and \bar{e}_j must satisfy the following:

$$\bar{e}_i \Delta \bar{e}_j = 0$$

This fact can be shown by observing that the symmetric matrix $\Delta^{1/2}P\Delta^{-1/2}$ has orthonormal eigenvectors. Furthermore, if \bar{x} is an eigenvector of P , then $\Delta^{1/2}\bar{x}$ is an eigenvector of the symmetric matrix. Since any pair of eigenvectors $\Delta^{1/2}\bar{x}_i$ and $\Delta^{1/2}\bar{x}_j$ of the symmetric matrix are orthonormal, it follows that $\bar{x}_i^T \Delta \bar{x}_j = 0$. The result follows.

4. Show that all eigenvectors (other than the first eigenvector) of the stochastic transition matrix of a connected, undirected graph will have both positive and negative components.

The top eigenvector of the stochastic transition matrix P is a vector of 1s. Let $\Lambda = \Delta^{1/2}$ be the square-root of the degree matrix. Then, the matrix $S = \Lambda P \Lambda^{-1}$ is a symmetric matrix in which the first eigenvector is $\Lambda \bar{e}$, where \bar{e} is a vector of 1s. All other eigenvectors of S are orthonormal to this vector with only positive components and will therefore have both positive and negative components. Furthermore, if \bar{x} is an eigenvector of P (different from the first eigenvector), then $\Lambda \bar{x}$ can easily be shown to be an eigenvector of S . Since the latter has both positive and negative components, it follows that the former does too.

5. Consider the adjacency matrix A of an $n \times n$ undirected graph, which is also bipartite. In a bipartite graph, the n vertices can be divided into two vertex sets V_1 and V_2 of respectively n_1 and n_2 vertices, so that all edges occur between vertices of V_1 and vertices of V_2 . The adjacency matrix of such a graph always has the following form for an $n_1 \times n_2$ matrix B :

$$A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}$$

Even though A is symmetric, B might not be symmetric. Given the eigenvectors and eigenvalues of A , show how you can perform the SVD of B quickly.

Each eigenvector of A is an $(n_1 + n_2)$ -dimensional vector \bar{x} , in which the first n_1 components are \bar{x}_1 and the remaining n_2 components are \bar{x}_2 . Furthermore, the first n_1 components of $A\bar{x}$ are $B\bar{x}_2$, and the remaining n_2 components are $B^T\bar{x}_1$. Therefore, we have:

$$\begin{aligned} B\bar{x}_2 &\propto \bar{x}_1 \\ B^T\bar{x}_1 &\propto \bar{x}_2 \end{aligned}$$

One can use the above relationships to show that $B^T B \bar{x}_2 \propto \bar{x}_2$, and $B B^T \bar{x}_1 \propto \bar{x}_1$. In other words, \bar{x}_1 contains the left singular vectors, \bar{x}_2 contains the right singular vectors, and the square root of the eigenvalues of $B^T B$ or $B B^T$ contains the singular values. One can use these to construct the SVD.

6. A complete directed graph is defined on n vertices and it contains all $n(n-1)$ possible edges in both directions between each pair of vertices (other than self-loops). Each edge weight is 1.

- (a) Give a short reason why all eigenvalues must be real.
- (b) Give a short reason why the eigenvalues must sum to 0.
- (c) Show that this graph has one eigenvalue of $(n-1)$ and $(n-1)$ eigenvalues are -1 .

- (a) All eigenvalues are real because the matrix is symmetric.
- (b) The trace of the matrix is 0, since it has no self-loops. Therefore, the eigenvalues must sum to 0.
- (c) The matrix $(A+I)$ has rank 1, because all rows are identical. Therefore, eigenvalues -1 has multiplicity of $(n-1)$. It remains to find an eigenvalue λ so that $A - \lambda I$ is singular, which corresponds to the only possible remaining eigenvector with different eigenvalue. The matrix $A - (n-1)I$ has rows that sum to 0. Therefore, it is singular and must have a null space containing the remaining eigenvector.

7. A complete bipartite graph (see Exercise 5) is defined on 4 vertices, where 2 vertices are contained in each partition. A edge of weight 1 exists in both directions between each pair of vertices drawn from the two partitions. Find the eigenvalues of this graph. Can you generalize this result to the case of a complete bipartite graph containing $2n$ vertices, where n vertices are contained in each partition?

This matrix has 2 eigenvalues of 1, and $(2n - 2)$ eigenvalues of 0.

8. Suppose you create a symmetrically normalized adjacency matrix $S = \Delta^{-1/2} A \Delta^{-1/2}$ for an undirected adjacency matrix A . You decide that some vertices are “important” and they should get relative weight $\gamma > 1$ in an embedding that is similar to spectral clustering, whereas other vertices only get a weight to 1.
- (a) Propose a weighted matrix factorization model that creates an embedding in which the “important” vertices have a relative weight of γ in the objective function. The matrix factorization model should yield the same embedding at $\gamma = 1$ as symmetric spectral clustering.
 - (b) Show how you can create an informative embedding with this approach, if some vertices in the graph are labeled.
 - (c) You are given a black-box classifier that works with multidimensional data. Show how you can select γ appropriately and use it for collective classification of the unlabeled vertices of the graph.

The approach is a simple application of weighted matrix factorization methods discussed in Chapter 7. The only difference is that one is performing symmetric matrix factorization in this case. The approach can be used for semi-supervised embedding because it is possible to create a similarity graph and extract the embedding from it. Labeled vertices can be weighted to a greater degree.

9. Propose an embedding-based algorithm for outlier detection in multidimensional data that uses the concept of the similarity graph and the extraction of an embedding. Discuss the choice of an appropriate dimensionality of the embedding, and how this choice is different from the case of the clustering problem.

One can construct a similarity graph using any kernel similarity like the Gaussian kernel. Subsequently one can extract all nonzero eigenvectors, and perform whitening on the representation. The points with the largest distance from the centroid are outliers. This is simply the kernel Mahalanobis method discussed in Chapter 8.

11. Provide an example of a 2×2 adjacency matrix of a directed graph that is not diagonalizable.

The following graph containing a single edge between two vertices is not diagonalizable because it has an eigenvalue of 0, which has algebraic multiplicity of 2, and geometric multiplicity of 1:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

12. A bipartite graph is defined as a graph $G = (V_1 \cup V_2)$ with a partitioned vertex set $V_1 \cup V_2$, so that no edges in E exist within vertices of a single partition. In other words, for all $(i, j) \in E$, both i, j cannot be drawn from V_1 , and both i, j cannot be drawn from V_2 . Show that if λ is the eigenvalue of the adjacency matrix of an undirected bipartite graph, then $-\lambda$ is an eigenvalue as well.

Since the graph is bipartite, its adjacency matrix with appropriately reordered vertices is in the following form:

$$A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}$$

The eigenvector can be expressed in the form $[\bar{x}^T, \bar{y}^T]^T$, where $B\bar{y} = \lambda\bar{x}$, and $B^T\bar{x} = \lambda\bar{y}$. In such a case, $[-\bar{x}^T, \bar{y}^T]^T$ is an eigenvector with eigenvalue $-\lambda$.

- 15.** Let P be the stochastic transition matrix of an undirected and connected graph. Show that all left eigenvectors of P other than the principal left eigenvector (i.e., PageRank vector) have vector components that sum to 0. [Hint: What are the angles between left eigenvectors and right eigenvectors of a matrix?]

As discussed in Exercise 17 of Chapter 3, the left eigenvectors and right eigenvectors are orthogonal as long as they belong to distinct eigenvalues. The primary right eigenvectors of the transition matrix is a column of 1s. Therefore, all non-primary left eigenvectors will be orthogonal to a column of 1s. This is possible only when their components sum to 0.

- 16.** Let S be the symmetrically normalized adjacency matrix of spectral clustering. In some cases, the clusters do not clearly separate out by applying the k -means algorithm on the features obtained from eigenvector extraction on S . Use the kernel intuition from Chapter 8 to discuss the advantages of using $S \odot S$ instead of S for eigenvector extraction in such cases.

The k -means algorithm works well with clusters that are linearly separable. When this is not the case, the performance will be poor. One approach to improve performance is to apply a super-linear function on the entries, which increases the dimensionality of the embedding and promotes linear separability. A larger number of eigenvectors would need to be used for effective results.

- 17.** Consider two $n \times n$ symmetric matrices A and B , such that B is also positive definite. Show that BA need not be symmetric, but it is diagonalizable with real eigenvalues.

The fact that BA is not symmetric can be shown by example, by selecting B to be the weighted degree matrix of A , and selecting A to be the undirected 3×3 adjacency matrix of a cycle of three vertices in which one edge is 2 and other two edges are 1. Since B is positive definite but the square-root matrix $B^{1/2}$ and the inverse square root matrix $B^{-1/2}$ exist. The square-root matrix and inverse square-root matrix are both symmetric. Therefore, the matrix $C = B^{1/2}AB^{1/2}$ is symmetric, and has real eigenvectors and eigenvalues. Furthermore, the matrix $B^{1/2}CB^{-1/2}$ is similar to C and is equal to BA . Therefore C and BA have the same real eigenvalues.

- 18.** Suppose that A is the 20×20 adjacency matrix of a directed graph of 20 nodes. Interpret the matrix $(I - A^{20})(I - A)^{-1}$ in terms of walks in the graph. Will this matrix have any special properties for a strongly connected graph? Argue algebraically why the following is true:

$$(I - A^{20})(I - A)^{-1} = (I - A)^{-1}(I - A^{20})$$

This can be shown by expanding $(I - A^{20})$ as follows:

$$I - A^{20} = (I - A)(I + A + \dots + A^{19}) = (I + A + \dots + A^{19})(I - A)$$

This can be interpreted as all walks of length 19 or less in the graph. For a strongly connected graph, it will result in all entries of the matrix being positive.

19. Exercise 13 of the previous chapter introduces symmetric non-negative matrix factorization, which can also be used to factorize the symmetrically normalized adjacency matrix $S \approx UU^T$, which is used in spectral clustering. Here, U is an $n \times k$ non-negative factor matrix. Discuss why the top- r components of each column of U directly provide clustered bags of nodes of size r in the graph.

Note that outer product of a column with itself forms an important subgraph. NMF reconstructs the graph as an additive sum of these subgraphs. The top components of this subgraph form a relatively dense clique.

20. Find the PageRank of each node in (i) an undirected cycle of n nodes, and (ii) a single central node connected with an undirected edge to each of $(n - 1)$ nodes. In each case, compute the PageRank at a restart probability of 0.

The probability is $1/n$ for all nodes in the cycle. The probability is $1/2$ for the central node and $1/(2n - 2)$ for all other nodes.

21. **Signed network embedding:** Suppose that you have a graph with both positive and negative weights on edges. Propose modifications of the algorithms used to remove “weak edges” and to symmetrically normalize the graph. Will the resulting graph be diagonalizable with orthogonal eigenvectors and real eigenvalues? Is there anything special about the first eigenvector?

Weak edges can be removed based on the absolute magnitudes of the edges. Similarly, the absolute magnitudes of the edges are used to compute the sum of each row and column for symmetric normalization. The matrix is symmetric and therefore it does have orthogonal eigenvectors and real eigenvalues. There is nothing special about the first eigenvector because of the presence of both negative and positive edges.

22. **Heterogeneous network embedding:** Consider a social network graph with directed/undirected edges of multiple types (e.g., undirected friendship links, directed messaging links, and directed “like” links). Propose a shared matrix factorization algorithm to extract an embedding of each node. How would you tune the parameters?

Let us extract the adjacency matrices for each of the k different types of links and denote them by $A_1 \dots A_p$. Then, we perform the shared matrix factorization $A_i \approx UV_i^T$ for each $i \in \{1 \dots p\}$. Here, U is the shared factor across all link types, which provides the embedding of each node. On the other hand, V_i is the embedding of a node that captures its behavior with respect to a particular modality. The corresponding objective function is as follows:

$$J = \sum_i \beta_i \|A_i - UV_i^T\|^2$$

We set the error matrices to $E_i = A_i - UV_i^T$ for each i . Then, the updates are as follows:

$$\begin{aligned} U &\leftarrow U + \alpha \sum_i \beta_i E_i V \\ V_i &\leftarrow V_i + \alpha \beta_i E_i^T U \quad \forall i \end{aligned}$$

In the event that regularization is used, the updates are as follows:

$$U \Leftarrow U(1 - \alpha\lambda) + \alpha \sum_i \beta_i E_i V$$

$$V_i \Leftarrow V_i(1 - \alpha\lambda) + \alpha \beta_i E_i^T U \quad \forall i$$

The parameters are set by using out-of-sample root-mean-square-error (RMSE) of link weight prediction. Some of the edges are held out (along with negative edges of zero weight), and their RMSE is computed over different settings of the parameters. The choice of parameters with lowest error is selected.

Chapter 11

Optimization in Computational Graphs

1. One of the problems in this chapter proposes a loss function for the L_1 -SVM in the context of a computational graph. How would you change this loss function, so that the same computational graph results in an L_2 -SVM?

The new loss function is $\max\{0, 1 - o\hat{o}\}^2$, while using the same notations as those used in the book.

2. Repeat Exercise 1 with the changed setting that you want to simulate Widrow-Hoff learning with the same computational graph. What will be the loss function associated with the single output node? Widrow-Hoff learning is described in Chapter 3.

In this case, the loss function is $(1 - o\hat{o})^2$.

3. The book discusses a vector-centric view of backpropagation in which backpropagation in linear layers can be implemented with matrix-to-vector multiplications. Discuss how you can deal with batches of training instances at a time (i.e., batch stochastic gradient descent) by using matrix-to-matrix multiplications.

In the case of vector-centric back propagation, we stack up the training instances in columns to create a matrix. Subsequently, exactly the same operations are applied to the matrix in both the forward and backward phases, as were originally applied to the vector.

4. Let $f(x)$ be defined as follows:

$$f(x) = \sin(x) + \cos(x)$$

Consider the function $f(f(f(f(x))))$. Write this function in closed form to obtain an appreciation of the awkwardly long function. Evaluate the derivative of this function at $x = \pi/3$ radians by using a computational graph abstraction.

$$\begin{aligned} &\sin(\sin(\sin(\sin(x) + \cos(x)) + \cos(\sin(x) + \cos(x))) + \cos(\sin(\sin(x) + \cos(x)) + \\ &\cos(\sin(x) + \cos(x)))) + \\ &+ \cos(\sin(\sin(\sin(x) + \cos(x)) + \cos(\sin(x) + \cos(x))) + \cos(\sin(\sin(x) + \cos(x)) + \\ &\cos(\sin(x) + \cos(x)))) \end{aligned}$$

Construct the following computational graph of four nodes. A single input x feeds into a sine function node and a cosine function node. Both these nodes feed into a single addition node. This graph can be replicated four times in order to create the full computational graph. Then, the backpropagation algorithm can be applied in order to compute the derivative of the output with respect to the input.

5. Suppose that you have a computational graph with the constraint that specific sets of weights are always constrained to be at the same value. Discuss how you can compute the derivative of the loss function with respect to these weights.

We simply compute the derivative with respect to the shared weights separately and then we add them. This trick is also discussed at the end of section 3.4 in Chapter 3.

6. Consider a computational graph in which you are told that the variables on the edges satisfy k linear equality constraints. Discuss how you would train the weights of such a graph. How would your answer change, if the variables satisfied box constraints.

One can use row reduction in order to express a subset of the variables in terms of the others. Subsequently, the chain rule can be used in order to compute the derivatives with respect to the variables.

7. Discuss why the dynamic programming algorithm for computing the gradients will not work in the case where the computational graph contains cycles.

This is because one always needs to find a node in the forward phase in which the values of its incoming nodes are already defined. The same is true for the outgoing nodes in the backward phase. This is not possible in a graph with cycles, where an acyclic ordering of processing cannot be defined.

8. Consider the neural architecture with connections between alternate layers. Suppose that the recurrence equations of this neural network are as follows:

$$\begin{aligned}\bar{h}_1 &= \text{ReLU}(W_1 \bar{x}) \\ \bar{h}_2 &= \text{ReLU}(W_2 \bar{x} + W_3 \bar{h}_1) \\ y &= W_4 \bar{h}_2\end{aligned}$$

Here, W_1 , W_2 , W_3 , and W_4 are matrices of appropriate size. Use the vector-centric backpropagation algorithm to derive the expressions for $\frac{\partial y}{\partial \bar{h}_2}$, $\frac{\partial y}{\partial \bar{h}_1}$, and $\frac{\partial y}{\partial \bar{x}}$ in terms of the matrices and activation values in intermediate layers.

Using the linear layer backpropagation, we obtain the following:

$$\frac{\partial y}{\partial \bar{h}_2} = W_4^T \frac{\partial y}{\partial y} = W_4^T$$

Applying the vector-centric recurrence to the next layer, we obtain:

$$\frac{\partial y}{\partial \bar{h}_1} = \frac{\partial \bar{h}_2}{\partial \bar{h}_1} \frac{\partial y}{\partial \bar{h}_2} = W_3^T (I(\bar{h}_2 > 0) \odot W_4^T)$$

Here, $I(\cdot)$ is an element-wise indicator function, which takes on the value of 0 or 1 for each component, depending on whether or not the condition is satisfied for that

component. Finally, we can obtain an expression for $\frac{\partial y}{\partial \bar{x}}$ as follows:

$$\begin{aligned}\frac{\partial y}{\partial \bar{x}} &= \frac{\partial \bar{h}_2}{\partial \bar{x}} \frac{\partial y}{\partial \bar{h}_2} + \frac{\partial \bar{h}_1}{\partial \bar{x}} \frac{\partial y}{\partial \bar{h}_1} \\ &= W_2^T (I(\bar{h}_2 > 0) \odot W_4^T) + W_1^T \{I(\bar{h}_1 > 0) \odot [W_3^T (I(\bar{h}_2 > 0) \odot W_4^T)]\}\end{aligned}$$

9. Consider a neural network that has hidden layers $\bar{h}_1 \dots \bar{h}_t$, inputs $\bar{x}_1 \dots \bar{x}_t$ into each layer, and outputs \bar{o} from the final layer \bar{h}_t . The recurrence equation for the p th layer is as follows:

$$\begin{aligned}\bar{o} &= U \bar{h}_t \\ \bar{h}_p &= \tanh(W \bar{h}_{p-1} + V \bar{x}_p) \quad \forall p \in \{1 \dots t\}\end{aligned}$$

The vector output \bar{o} has dimensionality k , each \bar{h}_p has dimensionality m , and each \bar{x}_p has dimensionality d . The “tanh” function is applied in element-wise fashion. The notations U , V , and W are matrices of sizes $k \times m$, $m \times d$, and $m \times m$, respectively. The vector \bar{h}_0 is set to the zero vector. Start by drawing a (vectored) computational graph for this system. Show that node-to-node backpropagation uses the following recurrence:

$$\begin{aligned}\frac{\partial \bar{o}}{\partial \bar{h}_t} &= U^T \\ \frac{\partial \bar{o}}{\partial \bar{h}_{p-1}} &= W^T \Delta_{p-1} \frac{\partial \bar{o}}{\partial \bar{h}_p} \quad \forall p \in \{2 \dots t\}\end{aligned}$$

Here, Δ_p is a diagonal matrix in which the diagonal entries contain the components of the vector $1 - \bar{h}_p \odot \bar{h}_p$. What you have just derived contains the node-to-node backpropagation equations of a recurrent neural network. What is the size of each matrix $\frac{\partial \bar{o}}{\partial \bar{h}_p}$?

It is easy to show that the (i, j) entry of the first derivative is the (j, i) th entry of U because of the linear form of the update. For the second identity, one must use the chain rule over the tanh and the weight matrix product, which contribute Δ_{p-1} and W^T , respectively. Note that these matrices, which have k columns and m rows.

10. Show that if we use the loss function $L(\bar{o})$ in Exercise 9, then the loss-to-node gradient can be computed for the final layer \bar{h}_t as follows:

$$\frac{\partial L(\bar{o})}{\partial \bar{h}_t} = U^T \frac{\partial L(\bar{o})}{\partial \bar{o}}$$

The updates in earlier layers remain similar to Exercise 9, except that each \bar{o} is replaced by $L(\bar{o})$. What is the size of each matrix $\frac{\partial L(\bar{o})}{\partial \bar{h}_p}$?

This is similar to the previous exercise, except that we need to use the chain rule to append $\frac{\partial L(\bar{o})}{\partial \bar{o}}$ to the end (right side) of each matrix product as an additional factor. This change results in an m -dimensional column vector, because we are trying to find the gradient of the loss with respect to the hidden layer.

11. Suppose that the output structure of the neural network in Exercise 9 is changed so that there are k -dimensional outputs $\bar{o}_1 \dots \bar{o}_t$ in each layer, and the overall loss is

$L = \sum_{i=1}^t L(\bar{o}_i)$. The output recurrence is $\bar{o}_p = U\bar{h}_p$. All other recurrences remain the same. Show that the backpropagation recurrence of the hidden layers changes as follows:

$$\begin{aligned}\frac{\partial L}{\partial \bar{h}_t} &= U^T \frac{\partial L(\bar{o}_t)}{\partial \bar{o}_t} \\ \frac{\partial L}{\partial \bar{h}_{p-1}} &= W^T \Delta_{p-1} \frac{\partial L}{\partial \bar{h}_p} + U^T \frac{\partial L(\bar{o}_{p-1})}{\partial \bar{o}_{p-1}} \quad \forall p \in \{2 \dots t\}\end{aligned}$$

The gradient of the final layer \bar{h}_t is not change. However, for $p < t$, the layer \bar{h}_p accumulates the gradients of the loss with respect to \bar{h}_p .

12. For Exercise 11, show the following loss-to-weight derivatives:

$$\frac{\partial L}{\partial U} = \sum_{p=1}^t \frac{\partial L(\bar{o}_p)}{\partial \bar{o}_p} \bar{h}_p^T, \quad \frac{\partial L}{\partial W} = \sum_{p=2}^t \Delta_{p-1} \frac{\partial L}{\partial \bar{h}_p} \bar{h}_{p-1}^T, \quad \frac{\partial L}{\partial V} = \sum_{p=1}^t \Delta_p \frac{\partial L}{\partial \bar{h}_p} \bar{x}_p^T$$

What are the sizes and ranks of these matrices?

A key point is that the weights are shared across different layers. The first step is to pretend that the weights are not shared across layers and compute the derivative separately with respect to each weight matrix. Subsequently, the chain rule can be used to infer that these derivatives need to be added. Furthermore, the derivative in each layer can be obtained by multiplying Δ_{p-1} with $\frac{\partial L}{\partial \bar{h}_p}$ to obtain derivatives with respect to pre-activation values. Subsequently, loss to weight derivatives are obtained by taking the outer-product of the loss-to-preactivation derivative with \bar{h}_{p-1} (as discussed in the vector-centric backpropagation section of book).

13. Consider a neural network in which a vectored node \bar{v} feeds into two distinct vectored nodes \bar{h}_1 and \bar{h}_2 computing different functions. The functions computed at the nodes are $\bar{h}_1 = \text{ReLU}(W_1 \bar{v})$ and $\bar{h}_2 = \text{sigmoid}(W_2 \bar{v})$. We do not know anything about the values of the variables in other parts of the network, but we know that $\bar{h}_1 = [2, -1, 3]^T$ and $\bar{h}_2 = [0.2, 0.5, 0.3]^T$, that are connected to the node $\bar{v} = [2, 3, 5, 1]^T$. Furthermore, the loss gradients are $\frac{\partial L}{\partial \bar{h}_1} = [-2, 1, 4]^T$ and $\frac{\partial L}{\partial \bar{h}_2} = [1, 3, -2]^T$, respectively. Show that the backpropagated loss gradient $\frac{\partial L}{\partial \bar{v}}$ can be computed in terms of W_1 and W_2 as follows:

$$\frac{\partial L}{\partial \bar{v}} = W_1^T \begin{bmatrix} -2 \\ 0 \\ 4 \end{bmatrix} + W_2^T \begin{bmatrix} 0.16 \\ 0.75 \\ -0.42 \end{bmatrix}$$

What are the sizes of W_1 , W_2 , and $\frac{\partial L}{\partial \bar{v}}$?

First note that the sizes of W_1 and W_2 are both 3×4 , so that they can map a 4-dimensional vector to a 3-dimensional vector. The gradient of L with respect to \bar{v} is a 4-dimensional vectors, since \bar{v} is 4-dimensional. Using the multivariable chain rule for vectors, we obtain the following:

$$\frac{\partial L}{\partial \bar{v}} = \frac{\partial \bar{h}_1}{\partial \bar{v}} \frac{\partial L}{\partial \bar{h}_1} + \frac{\partial \bar{h}_2}{\partial \bar{v}} \frac{\partial L}{\partial \bar{h}_2}$$

Now note that because of the use of the combination of linear and ReLU to obtain \bar{h}_1 , the matrix $\frac{\partial \bar{h}_1}{\partial \bar{v}}$ is $W_1^T \Delta_1$, where Δ_1 is a diagonal matrix containing 1, 0, and 1 along diagonal entries. Multiplying the diagonal matrix with the given values of $\frac{\partial L}{\partial h_1}$ yields $[-2, 0, 4]^T$. Similarly, because of the combination of linear and sigmoid to obtain \bar{h}_2 , the matrix $\frac{\partial \bar{h}_2}{\partial \bar{v}}$ is the diagonal matrix containing the elements of $\bar{h}_2 \odot (1 - \bar{h}_2)$ on its diagonal entries, which is 0.16, 0.25, and 0.21. Multiplying the diagonal matrix with the given values of $\frac{\partial L}{\partial h_2}$ yields $[0.16, 0.75, -0.42]^T$.

- 14. Forward Mode Differentiation:** *The backpropagation algorithm needs to compute node-to-node derivatives of all nodes with respect to output nodes, and therefore computing gradients in the backwards direction makes sense. Consequently, the pseudocode in the chapter propagates gradients in the backward direction. However, consider the case where we want to compute the node-to-node derivatives of all nodes with respect to source (input) nodes $s_1 \dots s_k$. Propose a variation of the pseudocode in the book that computes node-to-node gradients in the forward direction. Why is the backward mode preferred for neural network training?*

The pseudocode of this algorithm is as follows:

```

Initialize  $S(s_r, s_r) = 1$  for all source nodes  $r \in \{1 \dots k\}$ ;
repeat
  Select an unprocessed node  $j$  such that the values of  $S(s_r, i)$  all of its incoming
  nodes  $i \in A_{in}(j)$  are available for all  $r$ ;
  Update  $S(s_r, j) \leftarrow \sum_{i \in A_{in}(j)} S(s_r, i) z(i, j)$  for all  $r$ ;
until all nodes have been selected;
```

The backward mode is preferred in neural network training because it is the loss (at the sink) that needs to be differentiated with respect to all nodes.

- 15. All-pairs node-to-node derivatives:** *Let $y(i)$ be the variable in node i in a directed acyclic computational graph containing n nodes and m edges. Consider the case where one wants to compute $S(i, j) = \frac{\partial y(j)}{\partial y(i)}$ for all pairs of nodes in a computational graph, so that at least one directed path exists from node i to node j . Propose an algorithm for all-pairs derivative computation that requires at most $O(n^2 m)$ time.*

For directly connected nodes we initialize $S(i, j, 1) = z(i, j)$ for all edges. All other $S(i, j, t)$ are set to 0. One scans each edge one by one. The recurrence relation is as follows for each edge (i, j) and each outgoing edge (j, k) of j :

$$S(i, k, t + 1) = S(i, k, t) + S(i, j, t) z(j, k)$$

One repeats this process until all paths of length up to $n - 1$ have been processed.

There is an alternative and equivalent way to do this by changing the recurrence. One can implement this recurrence by using the incoming edges of i rather than the outgoing edges of i as follows. For each incoming node k of i , one could perform the following update:

$$S(k, j, t + 1) = S(k, j, t) + S(i, j, t) z(k, i)$$

Finally, we add up $S(i, j, t)$ over different values of t to obtain $S(i, j)$.

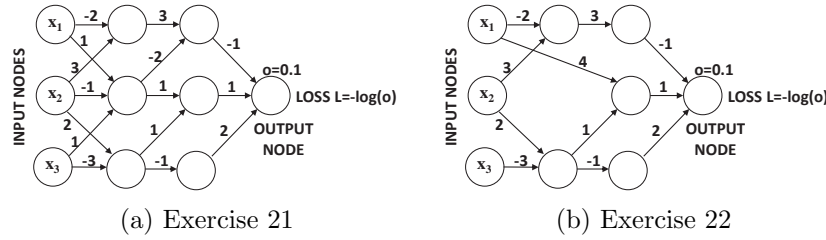


Figure 11.1: Computational graphs for Exercises 21 and 22

- 16.** Use the path-wise aggregation lemma to compute the derivative of $y(10)$ with respect to each of $y(1)$, $y(2)$, and $y(3)$ as an algebraic expression. You should get the same derivative as obtained using the backpropagation algorithm in the text of the chapter.

The paths are as follows:

From 1 to 10: 1, 4, 7, 10 and 1, 5, 8, 10

From 2 to 10: 2, 4, 7, 10, and 2, 5, 8, 10 and 2, 6, 9, 10

From 3 to 10: 3, 5, 8, 10, and 3, 6, 9, 10

Using these paths, the pathwise aggregation lemma yields the following expressions for $S(1, 10)$, $S(2, 10)$ and $S(3, 10)$:

$$S(1, 10) = z(1, 4)z(4, 7)z(7, 10) + z(1, 5)z(5, 8)z(8, 10)$$

$$S(2, 10) = z(2, 4)z(4, 7)z(7, 10) + z(2, 5)z(5, 8)z(8, 10) + z(2, 6)z(6, 9)z(9, 10)$$

$$S(3, 10) = z(3, 5)z(5, 8)z(8, 10) + z(3, 6)z(6, 9)z(9, 10)$$

We use the same notations $z(i, j)$ and $S(i, j)$ as in the example used in the text. Note that the values of $z(i, j)$ are already listed in the text. Using the listed values of $z(i, j)$ we obtain the following expressions:

$$\begin{aligned} S(1, 10) &= \frac{\partial y(10)}{\partial y(1)} = S(4, 10) \cdot z(1, 4) + S(5, 10) \cdot z(1, 5) \\ &= y(8) \cdot y(9) \cdot \cos[y(4)] \cdot y(2) - y(7) \cdot y(9) \cdot \sin[y(5)] \cdot y(2) \cdot y(3) \\ S(2, 10) &= \frac{\partial y(10)}{\partial y(2)} = S(4, 10) \cdot z(2, 4) + S(5, 10) \cdot z(2, 5) + S(6, 10) \cdot z(2, 6) \\ &= y(8) \cdot y(9) \cdot \cos[y(4)] \cdot y(1) - y(7) \cdot y(9) \cdot \sin[y(5)] \cdot y(1) \cdot y(3) + \\ &\quad + y(7) \cdot y(8) \cdot \cos[y(6)] \cdot w_2 \\ S(3, 10) &= \frac{\partial y(10)}{\partial y(3)} = S(5, 10) \cdot z(3, 5) + S(6, 10) \cdot z(3, 6) \\ &= -y(7) \cdot y(9) \cdot \sin[y(5)] \cdot y(1) \cdot y(2) + y(7) \cdot y(8) \cdot \cos[y(6)] \cdot w_3 \end{aligned}$$

These are exactly the same expressions as obtained using the backpropagation algorithm in the text.

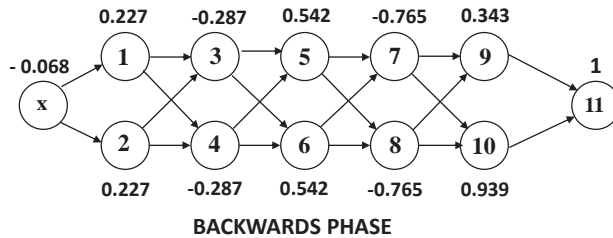
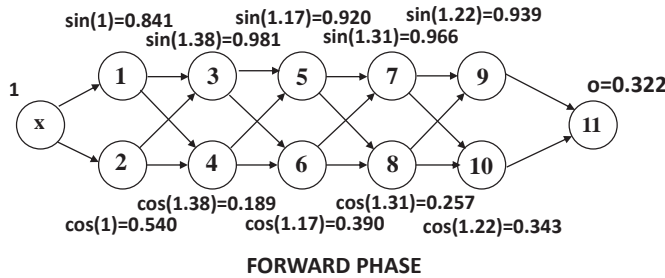
- 17.** Consider the computational graph of Figure 11.8. For a particular numerical input $x = a$, you find the unusual situation that the value $\frac{\partial y(j)}{\partial y(i)}$ is 0.3 for each and every edge (i, j) in the network. Compute the numerical value of the partial derivative of the

output with respect to the input x (at $x = a$). Show the computations using both the pathwise aggregation lemma and the backpropagation algorithm.

There are 32 paths of length 6 each. Each path contributes $0.3)^6$ to the derivative. Therefore, the total is 32×0.3^6 . One can show a similar result with the use of the backpropagation algorithm.

- 18.** Consider the computational graph of Figure 11.8. The upper node in each layer computes $\sin(x + y)$ and the lower node in each layer computes $\cos(x + y)$ with respect to its two inputs. For the first hidden layer, there is only a single input x , and therefore the values $\sin(x)$ and $\cos(x)$ are computed. The final output node computes the product of its two inputs. The single input x is 1 radian. Compute the numerical value of the partial derivative of the output with respect to the input x (at $x = 1$ radian). Show the computations using both the pathwise aggregation lemma and the backpropagation algorithm.

The forward phase activations as well as the derivatives in the backwards phase are shown in the figure below. It is important to note that all sine and cosine values are computed using radians rather than degrees:



- 19. Matrix factorization with neural networks:** Consider a neural network containing an input layer, a hidden layer, and an output layer. The number of outputs is equal to the number of inputs d . Each output value corresponds to an input value, and the loss function is the sum of squared differences between the outputs and their corresponding inputs. The number of nodes k in the hidden layer is much less than d . The d -dimensional rows of a data matrix D are fed one by one to train this neural network. Discuss why this model is identical to that of unconstrained matrix factorization of rank- k . Interpret the weights and the activations in the hidden layer in the context of matrix factorization. You may assume that the matrix D has full column rank. Define weight matrix and data matrix notations as convenient.

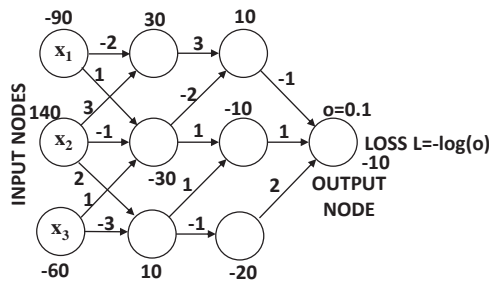
Let W_1 be the $d \times k$ weight matrix between the input and hidden layer. Let W_2 be the $k \times d$ weight matrix between the hidden and output layer. Let D be the $n \times d$ data matrix whose rows are fed to the neural network one by one. Then, the computational graph computes the matrix DW_1W_2 by passing the rows of D through the two layers. Therefore, the overall error is given by $\|D - DW_1W_2\|_F^2$. We can show the equivalence to unconstrained matrix factorization by means of a variable transformation. We simply use the new variables $U = DW_1$ and $V = W_2^T$ in order to show that the optimization problem is equivalent to minimizing $\|D - UV^T\|_F^2$. Note that this is the same optimization model as unconstrained matrix factorization. Furthermore, since D has full column rank, one can always find a W_1 exactly satisfying $DW_1 = U$ once U has been optimized.

- 20. SVD with neural networks:** *In the previous exercise, unconstrained matrix factorization finds the same k -dimensional subspace as SVD. However, it does not find an orthonormal basis in general like SVD. Provide an iterative training method for the computational graph of the previous section by gradually increasing the value of k so that an orthonormal basis is found.*

The training is performed by first performing the learning for $k = 1$, then fixing the weights for what has already been learned. Then, we perform the learning at $k = 2$, while training only the new weights. This process is repeated until the desired value of k has been reached. By training, the weights iteratively in this fashion, we ensure that the k -dimensional subspace found is the same as SVD is for *any* k less than the desired threshold.

- 21.** *Consider the computational graph shown in Figure 11.1(a), in which the local derivative $\frac{\partial y(j)}{\partial y(i)}$ is shown for each edge (i, j) , where $y(k)$ denotes the activation of node k . The output o is 0.1, and the loss L is given by $-\log(o)$. Compute the value of $\frac{\partial L}{\partial x_i}$ for each input x_i using both the path-wise aggregation lemma, and the backpropagation algorithm.*

The derivatives with respect to each node are shown in the diagram below. The backpropagated values are shown along with each node. On the right-hand side, the pathwise values are also shown.



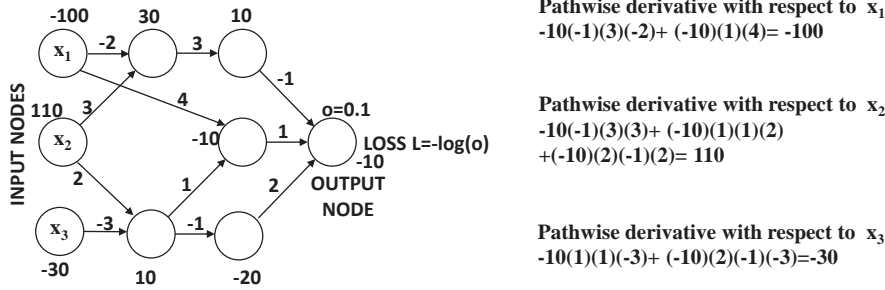
Pathwise derivative with respect to x_1
 $-10(-1)(3)(-2) + (-10)(-1)(-2)(1)$
 $+(-10)(1)(1)(1) = -90$

Pathwise derivative with respect to x_2
 $-10(-1)(3)(3) + (-10)(-1)(-2)(-1)$
 $+(-10)(1)(1)(-1) + (-10)(1)(1)(2)$
 $+(-10)(2)(-1)(2) = 140$

Pathwise derivative with respect to x_3
 $-10(-1)(-2)(1) + (-10)(1)(1)(1)$
 $+(-10)(1)(1)(-3) + (-10)(2)(-1)(-3) = -60$

- 22.** *Consider the computational graph shown in Figure 11.1(b), in which the local derivative $\frac{\partial y(j)}{\partial y(i)}$ is shown for each edge (i, j) , where $y(k)$ denotes the activation of node k . The output o is 0.1, and the loss L is given by $-\log(o)$. Compute the value of $\frac{\partial L}{\partial x_i}$ for each input x_i using both the path-wise aggregation lemma, and the backpropagation algorithm.*

The derivatives with respect to each node are shown in the diagram below. The backpropagated values are shown along with each node. On the right-hand side, the pathwise values are also shown.



23. Convert the weighted computational graph of linear regression into an unweighted graph by defining additional nodes containing $w_1 \dots w_5$ along with appropriately defined hidden nodes.

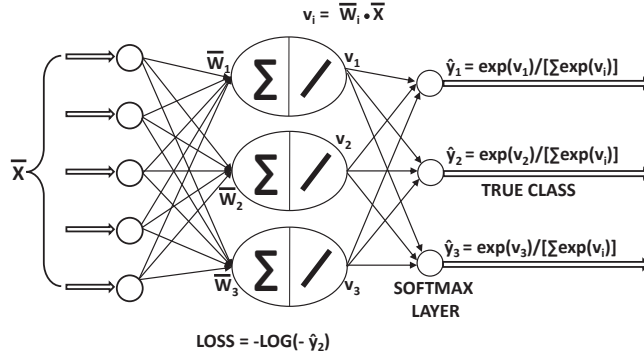
We need to add 5 input nodes containing $w_1 \dots w_5$, and five hidden nodes containing the multiplication function, creating a new hidden layer. Each node in the hidden layer computes $w_i * x_i$ for all i . Subsequently, the output node adds these five values.

24. **Multinomial logistic regression with neural networks:** Propose a neural network architecture using the softmax activation function and an appropriate loss function that can perform multinomial logistic regression.

It assumed that the input to the model is a training data set containing pairs of the form $(\bar{X}_i, c(i))$, where $c(i) \in \{1 \dots k\}$ is the index of the class of d -dimensional row vector \bar{X}_i . As in the case of the previous two models, the class r with the largest value of $\bar{W}_r \cdot \bar{X}_i^T$ is predicted to be the label of the data point \bar{X}_i . There is an additional probabilistic interpretation of $\bar{W}_r \cdot \bar{X}_i^T$ in terms of the posterior probability $P(r|\bar{X}_i)$ that the class r is predicted given the data point \bar{X}_i . This estimation can be naturally accomplished with the softmax activation function:

$$P(r|\bar{X}_i) = \frac{\exp(\bar{W}_r \cdot \bar{X}_i^T)}{\sum_{j=1}^k \exp(\bar{W}_j \cdot \bar{X}_i^T)} \quad (11.1)$$

Note that large values of $\bar{W}_r \cdot \bar{X}_i^T$ map to large probabilities, and the probabilities over all classes always sum to 1. The loss function L_i for the i th training instance is defined by the cross-entropy, which is the negative logarithm of the probability of the true class. The neural architecture of the softmax classifier is illustrated below:



The cross-entropy loss may be expressed in terms of either the input features or in terms of the softmax pre-activation values $v_r = \bar{W}_r \cdot \bar{X}_i^T$ as follows:

$$L_i = -\log[P(c(i)|\bar{X}_i)] \quad (11.2)$$

$$= -\bar{W}_{c(i)} \cdot \bar{X}_i^T + \log\left[\sum_{j=1}^k \exp(\bar{W}_j \cdot \bar{X}_i^T)\right] \quad (11.3)$$

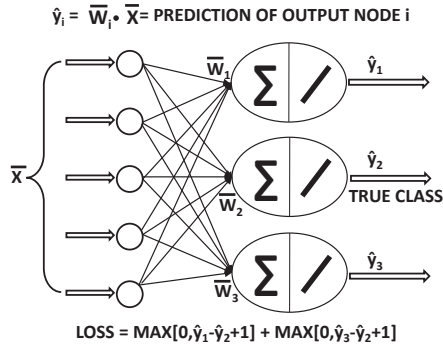
$$= -v_{c(i)} + \log\left[\sum_{j=1}^k \exp(v_j)\right] \quad (11.4)$$

25. Weston-Watkins SVM with neural networks: *Propose a neural network architecture and an appropriate loss function that is equivalent to the Weston-Watkins SVM.*

it is assumed that the i th training instance is denoted by $(\bar{X}_i, c(i))$, where \bar{X}_i contains the d -dimensional feature variables, and $c(i)$ contains the class index drawn from $\{1, \dots, k\}$. One wants to learn d -dimensional coefficients $\bar{W}_1 \dots \bar{W}_k$ of the k linear separators so that the class index r with the largest value of $\bar{W}_r \cdot \bar{X}_i^T$ is predicted to be the correct class $c(i)$. The Weston-Watkins loss function L_i for the i th training instance $(\bar{X}_i, c(i))$ is defined as follows:

$$L_i = \sum_{r:r \neq c(i)} \max(\bar{W}_r \cdot \bar{X}_i^T - \bar{W}_{c(i)} \cdot \bar{X}_i^T + 1, 0) \quad (11.5)$$

The neural architecture of the Weston-Watkins SVM is illustrated below:



Linear Algebra and Optimization for Machine Learning

A Textbook

Aggarwal, C.

2020, XXI, 495 p. 93 illus., 26 illus. in color., Hardcover

ISBN: 978-3-030-40343-0