Mathematics for Machine Learning
Problem Set 2
COMS 4995-002 : Summer B 2024

Instructor: Samuel Deng
samdeng@cs.columbia.edu
Due: 07/18/2024 11:59 PM ET.

Problem 1

Interpretation of the linear transformation under SVD (25 points total). You proved in Problem Set 1 that linear transformations are equivalent to matrices and vice versa. This was the following theorem:

(a) Any matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ defines a linear transformation $T : \mathbb{R}^d \to \mathbb{R}^n$ through matrix-vector multiplication:

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

(b) Any linear transformation $T : \mathbb{R}^d \to \mathbb{R}^n$ is given by matrix-vector multiplication by a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$:

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x},$$

where the ith column of $\mathbf{A}$ is $T(\mathbf{e}_i)$.

In this problem, we will use the *singular value decomposition (SVD)* of a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ to interpret how it transforms vectors on the *unit sphere*, the set of all unit vectors in $\mathbb{R}^n$. Keep in the back of your mind that multiplying any vector by the matrix $\mathbf{A}$ is the same as applying its associated linear transformation T. Recall from lecture that a *unit vector* is a vector $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{x}\| = 1$. We will denote the unit sphere as $\mathcal{S}$, the set of all such unit vectors:

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}.$$

In $\mathbb{R}^2$, the unit sphere is given by the all vectors $\mathbf{x} = (x_1, x_2)$ that satisfy the equation of a circle with radius 1:

$$\|\mathbf{x}\| = 1 \iff x_1^2 + x_2^2 = 1.$$

For the rest of this problem, let $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ be a matrix with $\text{rank}(\mathbf{A}) = 2$. Let $T_{\mathbf{A}} : \mathbb{R}^2 \to \mathbb{R}^2$ refer to its associated linear transformation. By the singular value decomposition,

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{u}_1, \mathbf{u}_2$ are the left singular vectors and columns of $\mathbf{U}$, $\mathbf{v}_1, \mathbf{v}_2$ are the right singular vectors and columns of $\mathbf{V}$, and $\Sigma \in \mathbb{R}^{2 \times 2}$ is a diagonal matrix with entries σ_1, σ_2, the singular values of $\mathbf{A}$.

Problem 1(a) [5 points] Prove that $T_{\mathbf{A}}(\mathbf{v}_1) = \sigma_1 \mathbf{u}_1$ and $T_{\mathbf{A}}(\mathbf{v}_2) = \sigma_2 \mathbf{u}_2$.

Problem 1(a) tells us that the transformation maps the right singular vectors to scaled left singular vectors. However, we want to interpret how $T_{\mathbf{A}}$ acts on any $\mathbf{x} \in \mathbb{R}^2$. It turns out that we can interpret this in terms of the left singular vectors, right singular vectors, and singular values.

Problem 1(b) [5 points] We can write $\mathbf{x}$ as:

$$\mathbf{x} = \nu_1 \mathbf{v}_1 + \nu_2 \mathbf{v}_2,$$

for unique scalars $\nu_1, \nu_2 \in \mathbb{R}$. State why these scalars are unique. Then, show that, for any $\mathbf{x} \in \mathbb{R}^2$,

$$T_{\mathbf{A}}(\mathbf{x}) = \nu_1 \sigma_1 \mathbf{u}_1 + \nu_2 \sigma_2 \mathbf{u}_2.$$

Hint: Problem 1(a) is helpful here.

Problem 1(b) gives us an expression that interprets the transformation of any vector $\mathbf{x} \in \mathbb{R}^2$ in terms of the left singular vectors of $\mathbf{A}$. It isn't completely clear yet how to interpret this, but one property of the left singular vectors that we can exploit is, again, that they form an orthonormal basis.

Recall from Problem 3 on Problem Set 1 that, given any basis for $\mathbb{R}^n$, we can write any vector $\mathbf{x} \in \mathbb{R}^n$ as a linear combination of the basis vectors. Its coordinates in that basis are the coefficients of that linear combination. Denote $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2)$ the ordered basis of right singular vectors. In this case, in Problem 1(b), the coordinates of $\mathbf{x}$ in the basis $\mathcal{V}$ are $[\mathbf{x}]_{\mathcal{V}} = (\nu_1, \nu_2)$. In the "language" of the basis $\mathcal{V}$, the vector $\mathbf{x}$ has coordinates (ν_1, ν_2). We will also use the "language" of the basis $\mathcal{U} := (\mathbf{u}_1, \mathbf{u}_2)$ to reinterpret $\mathbf{y} = T_{\mathbf{A}}(\mathbf{x})$.

Problem 1(c) [5 points] Let $\mathbf{y} = T_{\mathbf{A}}(\mathbf{x})$. We can write $\mathbf{y}$ as:

$$\mathbf{y} = \mu_1 \mathbf{u}_1 + \mu_2 \mathbf{u}_2,$$

for unique scalars $\mu_1, \mu_2 \in \mathbb{R}$. State why these scalars are unique. Let $[\mathbf{y}]_{\mathcal{U}} = (\mu_1, \mu_2)$. Then, show that, for any $\mathbf{x} \in \mathbb{R}^2$,

$$[\mathbf{y}]_{\mathcal{U}} = \mathbf{\Sigma} [\mathbf{x}]_{\mathcal{V}}.$$

Writing out Problem 1(c) explicitly, we have:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}. \tag{1}$$

Seeing it written like this, we can interpret the transformation $T_{\mathbf{A}}$ as follows. If we view $\mathbf{x}$, the vector that's being acted upon, in terms of the ordered basis $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2)$ and $\mathbf{y}$, the resulting vector, in terms of the ordered basis $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2)$, the transformation $T_{\mathbf{A}}$ looks as simple as it can get — a diagonal matrix!

Further, observe that Equation (1) immediately gives us the relations:

$$\nu_1 = \frac{\mu_1}{\sigma_1} \quad \text{and} \quad \nu_2 = \frac{\mu_2}{\sigma_2}. \tag{2}$$

So far, this problem has applied to any general $\mathbf{x} \in \mathbb{R}^2$. The motivation for this problem was to see what $T_{\mathbf{A}}$ specifically does to the vectors on the unit circle, $\mathbf{x} \in \mathcal{S}$.

Problem 1(d) [5 points]. Let $\mathbf{x} \in \mathcal{S}$ be a unit vector. Using the linear combination of $\mathbf{x}$ in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$ in Problem 1(b), prove that:

$$\nu_1^2 + \nu_2^2 = 1.$$

Conclude that

$$\frac{\mu_1^2}{\sigma_1^2} + \frac{\mu_2^2}{\sigma_2^2} = 1.$$

The final statement in Problem 1(d) is the equation of an elipse with a major axis $\mathbf{u}_1$ and minor axis $\mathbf{u}_2$. Using the SVD, you just proved that we can interpret the transformation of any full-rank matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ acting on the unit circle in $\mathbb{R}^2$ as mapping vectors onto an ellipse with axes characterized by the left singular vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ and scale characterized by the singular values σ_1^2 and σ_1^2. A larger singular value σ_1, for example, results in a "wider" axis in the $\mathbf{u}_1$ direction.

We won't prove this here (*optional:* you can try it yourself!), but the argument in this problem can be generalized to general matrices $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $\text{rank}(\mathbf{A}) = r$. In that case, we can prove that $T_{\mathbf{A}} : \mathbb{R}^d \to \mathbb{R}^n$ maps vectors in $\text{span}(\text{col}(\mathbf{A}^\top))$ onto an ellipsoid in $\text{span}(\text{col}(\mathbf{A}))$ with axes specified by the first r left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$. If $r \leq n$, then this is a "degenerate ellipsoid" in $\mathbb{R}^n$.

Finally, let us connect this back to eigendecomposition and diagonalization. Because the $\mathbf{A} \in \mathbb{R}^{2\times 2}$ we focused on in this problem is square, we might ask if we can interpret the transformation $T_{\mathbf{A}}$ in terms of its eigendecomposition. If $\mathbf{A}$ has the eigendecomposition

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top,$$

it is relatively straightforward to apply the same arguments we walked through in this problem to get an equation of the elipse in terms of the eigenvalues and eigenvectors.

Problem 1(e) [5 points] Suppose that $\mathbf{A} \in \mathbb{R}^2$ is a positive definite matrix, with eigendecomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$. Let $T_{\mathbf{A}}$ be the associated linear transformation. If its eigenvectors are $\mathbf{v}_1, \mathbf{v}_2$ and its eigenvalues are λ_1, λ_2, show that there exist unique scalars ν_1, ν_2 such that, for any $\mathbf{y} = T_{\mathbf{A}}(\mathbf{x}) \in \mathbb{R}^2$:

$$\mathbf{y} = \nu_1\mathbf{v}_1 + \nu_2\mathbf{v}_2.$$

Conclude, using the arguments above, that for any $\mathbf{x} \in \mathcal{S}$, the associated transformation $T_{\mathbf{A}}(\mathbf{x})$ maps vectors on the unit circle to the elipse defined by:

$$\frac{\nu_1^2}{\lambda_1} + \frac{\nu_2^2}{\lambda_2} = 1.$$

You may use any of the results you've already proven in Problem 1 here without proof.

Problem 2

Quadratic forms and positive semidefinite matrices (25 points total).

As briefly stated in lecture, a prevalent theme in this course is that a common technique to solve tough problems will be to look at nonlinear functions as approximated by linear functions. We have all the tools of linear algebra at our disposal when dealing with linear functions because of their equivalence with matrices, as we explored in Problem 1 of Problem Set 1.

As a next step up in complexity, we can look at *quadratic functions*, which describe a larger class of phenomena.[1] A general *quadratic function* $f : \mathbb{R}^d \to \mathbb{R}$ is a polynomial with degree two with d variables. For example, a single variable quadratic function $f : \mathbb{R} \to \mathbb{R}$ looks like:

$$f(x) = ax^2 + bx + c, \quad \text{where} \quad a \neq 0.$$

A two-variable quadratic function $f : \mathbb{R}^2 \to \mathbb{R}$ looks like:

$$f(x_1, x_2) = \underbrace{ax_1^2 + bx_1x_2 + cx_2^2}_{\text{“quadratic part”}} + \underbrace{dx_1 + ex_2}_{\text{“linear part”}} + \underbrace{f}_{\text{“constant part”}},$$

where a, b, or c is nonzero. In general, a quadratic function can have as many variables as you'd like, but they quickly start becoming unwieldy to write down explicitly. Note from the two equations above that all quadratic functions have a “quadratic part,” “a linear part,” and a “constant part.” We will focus on the “quadratic part” of quadratic functions, as that ends up dominating the shape and behavior of such functions.

The “quadratic part” of a quadratic function is called a quadratic form. Recall from lecture that a *quadratic form* is a polynomial function $f : \mathbb{R}^n \to \mathbb{R}$ with terms all of degree two. Some examples of quadratic forms $f : \mathbb{R}^2 \to \mathbb{R}$ include $f(x_1, x_2) = x_1^2 + x_2^2$ or $f(x_1, x_2) = 4x_1^2 + x_1x_2 - x_2^2$.

First, we will study quadratic forms using pure high school algebra. Recall the technique of “completing the square.” One can complete the square to prove a familiar formula from high school: the quadratic formula.

Problem 2(a) [3 points] Recall the quadratic formula: to find roots of single variable quadratic equations of the form $ax^2 + bx + c = 0$, it suffices to let

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

[1] In both physics and machine learning, we usually reach for linear functions to describe the “fist-order” effects of nonlinear phenomena. We reach for quadratic functions to describe the “second-order” effects. Usually, understanding the first and second order effects gives us a very good understanding on how a nonlinear function behaves. This will become more formal after our lecture on Taylor series.

Prove this formula by completing the square. *Hint:* add and subtract $\left(\frac{b}{2\sqrt{a}}\right)^2$, factor to complete the square, and solve for x.

Here's an important fact about quadratic forms that we will state here but not prove. For any quadratic form $f : \mathbb{R}^d \to \mathbb{R}$, there exist $r = p + l$ linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ such that

$$f(\mathbf{x}) = (\mathbf{v}_1^\top \mathbf{x})^2 + \cdots + (\mathbf{v}_p^\top \mathbf{x})^2 - (\mathbf{v}_{p+1}^\top \mathbf{x})^2 - \cdots - (\mathbf{v}_{p+l}^\top \mathbf{x})^2, \tag{3}$$

where p is the number of squares with positive coefficients and l is the number of squares with negative coefficients. This is, roughly, because one can always complete the square for a quadratic form.

There are certain quadratic forms that are quite nice to analyze. We say that a quadratic form $f : \mathbb{R}^d \to \mathbb{R}$ is *positive definite* if $f(\mathbf{x}) > 0$ when $\mathbf{x} \neq \mathbf{0}$. Note that, on its own, this property doesn't seem to have anything to do with the notion of positive definite matrices we learned in lecture, but the connection will soon become clear.

Problem 2(b) [3 points] Consider the following quadratic form $f(x_1, x_2) = x_1^2 + x_1 x_2$. Find linearly independent vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ such that $f(x_1, x_2)$ can be expressed as in Equation (3). In this case, what is p and what is l?

There is a term for these inner products in Equation (3). If we view each vector $\mathbf{v}_j$ for $j \in [n]$ as a matrix $\mathbf{v}_j^\top \in \mathbb{R}^{1 \times d}$, we can interpret, again, using our trusty theorem from Problem Set 1 that linear transformations are equivalent to matrices, the linear transformation that each vector is associated with. In general, for any vector $\mathbf{u} \in \mathbb{R}^d$, we can consider it as a linear function $T : \mathbb{R}^d \to \mathbb{R}$ if we view it as the row matrix $\mathbf{u} \in \mathbb{R}^{1 \times d}$; this is called a *linear functional* on $\mathbb{R}^d$. In this way, we can view any quadratic form as a sum of the squares of n linear functions on $\mathbb{R}^d$.

Problem 2(c) [3 points] Let $\mathbf{u} = (1, 2, 3)$. Using the equivalence of matrices and linear transformations theorem from Problem Set 1 (restated on the first page of this problem set), write the linear functional $T : \mathbb{R}^3 \to \mathbb{R}$ determined by $\mathbf{u}$.

Consider the set of vectors $\mathbf{x} \in \mathbb{R}^3$ such that $T(\mathbf{x}) = 0$. Prove that this set is a subspace, and find a basis for this subspace. What is the dimension of this subspace?

Just for yourself: what does this subspace look like?

It can be shown that for any quadratic, the numbers p and l are characteristic of the quadratic form — that is, regardless of how we write the quadratic form as a sum of squares (as in Equation (3)), the number p of plus signs and the number l of minus signs will always be the same for the same quadratic form. Because of this, the ordered pair (p, l) is often called the quadratic form's *signature* or *type*. Interestingly, this inherent property of every quadratic

form is linked to linear algebra; particularly, it is linked to the eigenvalues and eigenvectors of a matrix associated with every quadratic form.

In lecture, we saw that every two-variable quadratic form $f : \mathbb{R}^2 \to \mathbb{R}$ can be written in terms of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}.$$

It turns out that this is a general fact for quadratic forms $f : \mathbb{R}^d \to \mathbb{R}$.

(a) Any symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ defines an associated quadratic form $f : \mathbb{R}^d \to \mathbb{R}$ through

$$f_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}.$$

(b) Any quadratic form $f : \mathbb{R}^d \to \mathbb{R}$ has a *unique* associated symmetric matrix $\mathbf{A}$ such that

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}.$$

Problem 2(d) [3 points] Using constants a, b, c, d, e, and f in $\mathbb{R}$, write a general quadratic form $f : \mathbb{R}^3 \to \mathbb{R}$ in three variables. Write the quadratic form as $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{3\times 3}$ is a symmetric matrix with entries a, b, c, d, e and f.

For the specific quadratic form

$$f(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_1x_3 + 4x_2x_3 + 9x_3^2,$$

write the associated symmetric matrix $\mathbf{A} \in \mathbb{R}^{3\times 3}$.

Recall that a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is *positive definite* if

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}.$$

The connection in terminology to positive definite quadratic forms should now be apparent. One particularly important property about positive definite quadratic forms comes from the simple fact that we can always construct a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ from an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ by considering $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$.

Problem 2(e) [3 points] Verify that, for any $\mathbf{X} \in \mathbb{R}^{n\times d}$, the matrix $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$ is a symmetric matrix. Prove that the quadratic form defined by $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ is a positive definite quadratic form if and only if $\text{rank}(\mathbf{X}) = d$.

Because we can now study quadratic forms in terms of matrices, it should become clearer that the tools of linear algebra are at our disposal. Specifically, because quadratic forms are related to *symmetric* matrices, we might hope to use the spectral theorem and the eigenvectors and eigenvalues of $\mathbf{A}$ to analyze quadratic forms.

It turns out that we can do exactly that, and the eigenvectors and eigenvalues of $\mathbf{A}$ completely characterize its associated quadratic form. Without the tools of linear algebra, we characterized a quadratic form by its signature (p, l) via completing the square. With linear algebra, we take a different perspective of a quadratic form's signature through the eigenvalues and eigenvectors of $\mathbf{A}$, the associated symmetric matrix.

First, we prove a useful and important general property about orthonormal bases.

Problem 2(f) [3 points] Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^n$ be an orthonormal basis. Prove that any vector $\mathbf{x} \in \mathbb{R}^n$ can be written as

$$\mathbf{x} = \sum_{i=1}^{n} (\mathbf{x}^\top \mathbf{v}_i)\mathbf{v}_i.$$

Recall that the *spectral theorem* says that any symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ has the eigendecomposition:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top,$$

where $\mathbf{V} \in \mathbb{R}^{d \times d}$ is a matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_d$, the eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda} \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entries $\lambda_1, \ldots, \lambda_d$, the eigenvalues of $\mathbf{A}$.

Problem 2(g) [3 points] Using the spectral theorem and Problem 2(f), show that for any vector $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^\top \mathbf{v}_i)\mathbf{v}_i.$$

Finally, we will link the notion of a quadratic form's signature to the eigenvalues and eigenvectors. It turns out that the number of positive eigenvalues is exactly p, the number of negative eigenvalues is exactly l, and there are exactly $d - p - l$ zero eigenvalues.

Problem 2(h) [4 points] Prove that a quadratic form $f : \mathbb{R}^d \to \mathbb{R}$ with associated symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ has signature (p, l) if and only if there exists an orthonormal basis of eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ such that $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$, with $\lambda_1, \ldots, \lambda_p > 0$, $\lambda_{p+1}, \ldots, \lambda_{p+l} < 0$, and all of $\lambda_{p+l+1}, \ldots, \lambda_d$ are 0 (if $p + l < d$).

Hint: It suffices to show that $f(\mathbf{x}) = \sum_{i=1}^{d} \lambda_i (\mathbf{v}_i^\top \mathbf{x})^2$. Use Problem 2(f) and Problem 2(g) to obtain this equality.

When initially studied, quadratic forms seem like purely algebraic objects. Problem 2(h) shows us, however, that the tools of linear algebra allow us to interpret them exactly in terms of eigenvalues and eigenvectors. It tells us that everything we need to know about a quadratic form; for instance, whether its associated quadratic function "curves up" (is positive definite) can be read off from the eigenvalues of its associated symmetric matrix $\mathbf{A}$.

Problem 3

The pseudoinverse and errors in least squares regression (25 points total).

Recall the setup of *least squares regression* from lecture. We are given n training samples with d features $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ and a vector of training labels $\mathbf{y} \in \mathbb{R}^n$. Arranged row-wise, the training samples form a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with columns $\mathbf{x}_1, \ldots, \mathbf{x}_d$. For each $i \in [n]$, our goal is to make a prediction $\hat{y}_i \in \mathbb{R}$, such that $\hat{y}_i$ is as close to y_i as possible. In order to make these predictions, we need to construct a weight vector $\mathbf{w} \in \mathbb{R}^d$ such that $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i$. In matrix-vector form, we want to find $\mathbf{w} \in \mathbb{R}^d$ such that

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}} \approx \mathbf{y}.$$

The "least-squares" part of "least-squares regression" comes from how we model the approximation, "$\approx$." We want to find the $\mathbf{w} \in \mathbb{R}^d$ that minimizes a specific notion of error, the *sum of squared residuals* (also known as *mean squared error*), which we'll denote with $\text{err}(\cdot)$:

$$\text{err}(\mathbf{w}) := \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

In Problem Set 1, we investigated the errors of least squares regression just using our OLS solution for $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

In this problem, we will explore the *pseudoinverse* and its relationship to least squares regression and use the SVD to take a closer look at errors in least squares regression. Recall from lecture that, for any $\mathbf{X} \in \mathbb{R}^{n \times d}$ with full SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the *pseudoinverse* is defined as

$$\mathbf{X}^+ := \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^\top,$$

where $\mathbf{\Sigma}^+ = (\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^\top \in \mathbb{R}^{d \times n}$ if $n \geq d$ and $\mathbf{\Sigma}^+ = \mathbf{\Sigma}^\top(\mathbf{\Sigma}\mathbf{\Sigma}^\top)^{-1} \in \mathbb{R}^{d \times n}$ if $d > n$.

To get a grasp on this object, let's consider a concrete example first.

Problem 3(a) [3 points] Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Left singular vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \in \mathbb{R}^3$ for this matrix are given by:

$$\mathbf{u}_1 = \begin{bmatrix} \frac{5}{\sqrt{30}} \\ \frac{-2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{bmatrix} \quad \mathbf{u}_2 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \quad \mathbf{u}_3 = \begin{bmatrix} \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Right singular vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$ for this matrix are given by:

$$\mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{bmatrix} \quad \mathbf{v}_2 = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix}.$$

Finally, the singular values are $\sigma_1 = \sqrt{6}$ and $\sigma_2 = 1$. Write the matrix in full SVD form:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{3\times 3}$, $\mathbf{\Sigma} \in \mathbb{R}^{3\times 2}$ and $\mathbf{V} \in \mathbb{R}^{2\times 2}$. Compute the pseudoinverses $\mathbf{\Sigma}^+ \in \mathbb{R}^{2\times 3}$ and $\mathbf{A}^+ \in \mathbb{R}^{2\times 3}$. Verify numerically (showing your steps) that $\mathbf{\Sigma}^+\mathbf{\Sigma} = \mathbf{I}_{2\times 2}$. Also compute $\mathbf{\Sigma}\mathbf{\Sigma}^+$. You may use `numpy` or any other numerical computing software to compute these answers.

We already know from lecture that, when $\mathbf{X} \in \mathbb{R}^{n\times d}$, with $d \geq n$ and $\text{rank}(\mathbf{X}) = n$, using the pseudoinverse to obtain $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y}$ gives us the exact solution with smallest Euclidean norm. However, we've relied on the assumption that $\mathbf{X} \in \mathbb{R}^{n\times d}$ has full rank, i.e. $\text{rank}(\mathbf{X}) = \min\{n, d\}$, in all our uses of the pseudoinverse so far. This was necessary to invert the matrices $\mathbf{\Sigma}^\top\mathbf{\Sigma}$ or $\mathbf{\Sigma}\mathbf{\Sigma}^\top$.

Problem 3(b) [3 points] Consider $\mathbf{X} \in \mathbb{R}^{4\times 2}$ with singular values σ_1 and σ_2. Suppose that $\text{rank}(\mathbf{X}) = 2$. Compute $\mathbf{\Sigma}^\top\mathbf{\Sigma} \in \mathbb{R}^{2\times 2}$ and write it in terms of the singular values. Also, compute $\mathbf{\Sigma}^+ = (\mathbf{\Sigma}^\top\mathbf{\Sigma})^{-1}\mathbf{\Sigma}$ and write it in terms of the singular values.

Now, assume that $\text{rank}(\mathbf{X}) = 1$. Compute $\mathbf{\Sigma}^\top\mathbf{\Sigma} \in \mathbb{R}^{2\times 2}$ and write it in terms of the singular values. State why $\mathbf{\Sigma}^\top\mathbf{\Sigma}$ cannot be inverted.

There is an alternative characterization of the *pseudoinverse* using the *compact SVD* that will be easier to analyze for the purposes of this problem. Recall that any matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ has the compact SVD:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{V} \in \mathbb{R}^{d\times r}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$, where $r = \text{rank}(\mathbf{X})$. In this representation, the diagonal matrix $\mathbf{\Sigma}$ is only as large as the number of positive singular values $\sigma_1, \ldots, \sigma_r > 0$, which is equal to $\text{rank}(\mathbf{X})$. For the rest of this problem, we will focus on the pseudoinverse obtained from the compact SVD:

$$\mathbf{X}^+ = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top. \tag{4}$$

Because $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ is full-rank (its diagonal entries are all positive), there is no problem inverting $\mathbf{\Sigma}$ and this pseudoinverse is well-defined. We will now refer to Equation (4) as the pseudoinverse for the remainder of this problem, and we will consider the compact SVD unless stated otherwise.

Problem 3(c) [4 points] Prove these two properties of the pseudoinverse in Equation (4):

$$\mathbf{X}\mathbf{X}^+ = \mathbf{U}\mathbf{U}^\top \quad \text{and} \quad \mathbf{X}^+\mathbf{X} = \mathbf{V}\mathbf{V}^\top.$$

Also prove: (i) if $\text{rank}(\mathbf{X}) = n$, then $\mathbf{X}\mathbf{X}^\top = \mathbf{I}$ and (ii) if $\text{rank}(\mathbf{X}) = d$, then $\mathbf{X}^+\mathbf{X} = \mathbf{I}$.

Using the compact SVD also gives us the following simple property.

Problem 3(d) [3 points] Prove that, if $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ by the compact SVD,

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top,$$

where $\mathbf{V} \in \mathbb{R}^{d\times r}$ has columns $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

The compact SVD definition of the pseudoinverse gives us a solution to the normal equations even when $\mathbf{X}$ is not full rank. Recall that, to obtain a minimizer of $\text{err}(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, we solved the normal equations

$$(\mathbf{X}^\top\mathbf{X})\mathbf{w} = \mathbf{X}^\top\mathbf{y} \tag{5}$$

for $\mathbf{w} \in \mathbb{R}^d$. However, if $\text{rank}(\mathbf{X}) \leq d$, then the normal equations in Equation (5) may not have a unique solution.

Problem 3(e) [3 points] Using Problem 3(d), prove that, by using the pseudoinverse in Equation (4), the vector

$$\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} \in \mathbb{R}^d$$

is a solution to the normal equations.

Hint: Start by plugging in $\hat{\mathbf{w}}$ into the left hand side of Equation (5).

Problem 3(e) never assumed anything about the rank of $\mathbf{X} \in \mathbb{R}^{n\times d}$, so the pseudoinverse (of the compact SVD) given in Equation (4) has given us a solution that minimizes $\text{err}(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ without assuming that $\text{rank}(\mathbf{X}) = \min\{n, d\}$.

Another way to see this is as follows.

Problem 3(f) [3 points] Find the pseudoinverse of $\mathbf{X}^\top\mathbf{X}$, i.e. the matrix $(\mathbf{X}^\top\mathbf{X})^+ \in \mathbb{R}^{d\times d}$. Prove that

$$(\mathbf{X}^\top\mathbf{X})^+\mathbf{X}^\top\mathbf{y} = \mathbf{X}^+\mathbf{y}.$$

Problem 3(f) shows that our definition of pseudoinverse meshes well with our familiar solution for the normal equations. When $\text{rank}(\mathbf{X}) = d$, then $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. However, when $\text{rank}(\mathbf{X}) < d$, we can swap out $(\mathbf{X}^\top\mathbf{X})^{-1}$ with the pseudoinverse $(\mathbf{X}^\top\mathbf{X})^+$ and get a solution to the normal equations.

Finally, we will show that the pseudoinverse and the SVD give us a particularly illuminating perspective on the errors of our linear model, $\hat{\mathbf{w}} \in \mathbb{R}^d$. Recall from Problem Set 1 that a common way to model the errors we make from using a linear model is by positing that each sample has some error unexplained by the linear relationship, $\epsilon_i \in \mathbb{R}$. In this case, there exists some true underlying linear model $\mathbf{w}^* \in \mathbb{R}^d$, but the labels are now:

$$y_i = (\mathbf{w}^*)^\top \mathbf{x}_i + \epsilon_i. \tag{6}$$

We can collect all these errors into a vector, $\bar{\epsilon} \in \mathbb{R}^n$. As usual, we collect the true labels y_i into a vector $\mathbf{y} \in \mathbb{R}^n$. Writing Equation (6) with matrices and vectors, we get

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \bar{\epsilon}. \tag{7}$$

For the rest of this problem, assume that $\text{rank}(\mathbf{X}) = d$, for simplicity.

Problem 3(g) [3 points] Prove that with the error model in Equation (7), using the pseudoinverse characterization of the OLS solution in Problem 3(e),

$$\hat{\mathbf{w}} = \mathbf{X}^+ \mathbf{y},$$

the squared distance between $\hat{\mathbf{w}}$ and $\mathbf{w}^*$, the true linear model, satisfies

$$\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 = \bar{\epsilon}^\top \mathbf{V}\mathbf{\Sigma}^{-2}\mathbf{V}^\top \bar{\epsilon},$$

where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the compact SVD of $\mathbf{X}$.

Hint: First, show that $\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 = \|\mathbf{X}^+\bar{\epsilon}\|^2$. Problem 3(c) might be helpful. Then, expand $\|\mathbf{X}^+\bar{\epsilon}\|^2$ using the definition of the pseudoinverse.

What might $\bar{\epsilon}$ be, in the worst case? We don't have any notions of randomness right now (the third part of this course will cover such notions), but we can consider some "bad" choices of $\bar{\epsilon}$ that may occur. Particularly, we will consider what happens when $\bar{\epsilon}$ is in the direction of a right singular vector $\mathbf{v}_i$ with a small singular value σ_i.

Problem 3(h) [3 points] Consider any $i \in [r]$. Let $\bar{\epsilon} = \alpha \mathbf{v}_i$ where $\alpha \in \mathbb{R}$ is a scalar, and let $\sigma_i > 0$ be the singular value associated with $\mathbf{v}_i$. Prove, using Problem 3(g), that

$$\|\hat{\mathbf{w}} - \mathbf{w}^*\|^2 = \frac{\alpha}{\sigma_i^2}.$$

Problem 3(h) shows us that if σ_i is small, then the error in our estimate of $\mathbf{w}^*$ blows up, so long as $\bar{\epsilon}$ is in the direction of $\mathbf{v}_i$.

Programming Part

Eigendecomposition, PCA, and eigenfaces (25 points total). In this problem, you will use eigendecomposition to perform a basic dimensionality reduction technique in machine learning: principal components analysis (PCA).

In order to start this programming part, download the file `ps2.ipynb` from Course Content on the course webpage. Your submission for this part will be the same `ps2.ipynb` file modified with your code; see HW Submission on the course webpage for additional instructions.