## Homework 1

DUE: SATURDAY, JANUARY 25, 11:59PM

For the following problems, $m, n$ are assumed to be non-negative integers.

**Problem 1.1** (The Gram matrix $A^TA$)**.** Let $A \in \mathcal{M}_{m\times n}(\mathbb{R})$.

a) Show that $A^TA$ is an $n \times n$ symmetric matrix.
b) Show that $A^TA$ is positive semi-definite.
c) What do you need to assume about $A$ to ensure that $A^TA$ is positive definite, as opposed to just being positive semi-definite ?

The following might be useful:

$$M \in \mathcal{M}_{n\times n}(\mathbb{R}) \text{ is positive semi-definite} \quad \text{if and only if } M \text{ is symmetric and } \boldsymbol{x}^T M\boldsymbol{x} \geq 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^n \tag{1.1}$$

$$\text{if and only if } M \text{ is symmetric and has non-negative eigenvalues} \tag{1.2}$$

and

$$M \in \mathcal{M}_{n\times n}(\mathbb{R}) \text{ is positive definite} \quad \text{if and only if } M \text{ is symmetric and } \boldsymbol{x}^T M\boldsymbol{x} > 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \tag{1.3}$$

$$\text{if and only if } M \text{ is symmetric and has positive eigenvalues.} \tag{1.4}$$

Department of Mathematical Sciences, Carnegie Mellon University
*E-mail address*: jkoganem@andrew.cmu.edu.
*Date*: January 23, 2025.

**Problem 1.2.** Let

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{1.5}$$

a) Find the singular values of $A$.
b) Find a (full) singular-value decomposition of $A$.

**Problem 1.3** (Least squares and simple linear regression)**.** For this problem, we consider a dataset with $n$ samples $\{(x_i, y_i)\}_{i=1}^n$, and our goal is to model the underlying trend in the dataset by estimating the $i$-th sample via

$$y_i = \beta_0 + \beta_1 x_i. \tag{1.6}$$

In matrix form, this corresponds to solving the linear system $A\boldsymbol{x} = \boldsymbol{y}$, where

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \ \boldsymbol{x} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{1.7}$$

In general, $\boldsymbol{y} \notin \mathrm{Col}(A)$, since otherwise the linear model would capture the underlying trend perfectly. So instead, we look for the least squares solution $\hat{\boldsymbol{x}}$ satisfying

$$A^T A\hat{\boldsymbol{x}} = A^T \boldsymbol{y}. \tag{1.8}$$

This is the solution that makes $\boldsymbol{e} = \boldsymbol{b} - A\hat{\boldsymbol{x}} \in (\mathrm{Col}(A))^\perp = \mathrm{Null}(A^T)$, which is where (1.8) comes from. If the columns of $A$ are linearly independent, then we can write

$$\hat{\boldsymbol{x}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (A^T A)^{-1} A^T \boldsymbol{y}. \tag{1.9}$$

To make the algebra a little nicer, we introduce the quantities $\bar{x}, \bar{y}, \overline{x^2}, \overline{xy}, s_x, s_{xy}$ satisfying

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \ \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i, \ \overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2, \ \overline{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i$$
$$s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \ s_{xy} = \frac{1}{n}\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right) = \overline{xy} - \bar{x}\bar{y}. \tag{1.10}$$

In statistics, $\bar{x}, \bar{y}$ are referred to as the sample mean, $\frac{n}{n-1}s_x^2$ is referred to as the sample variance, and $\frac{n}{n-1}s_{xy}$ is referred to as the sample covariance (the use of $n-1$ instead of $n$ in the definitions of the sample variance and covariance is referred to as *Bessel's correction*). The square root of the sample variance is referred to as the sample standard deviation. The goal of this problem is to show that assuming $s_x^2 \neq 0$, we have

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{1.11}$$

a) Show that

$$\frac{1}{n}A^T A = \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}. \tag{1.12}$$

b) Use the formula for the inverse of $2 \times 2$ matrices,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \text{ if } ad - bc \neq 0, \tag{1.13}$$

to show that if $s_x^2 \neq 0$ then

$$\left(\frac{1}{n}A^T A\right)^{-1} = \frac{1}{s_x^2}\begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \tag{1.14}$$

c) Show that

$$\frac{1}{n}A^T \boldsymbol{y} = \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}. \tag{1.15}$$

d) Show that

$$\overline{x^2}\bar{y} - \bar{x}\overline{xy} = (s_x^2 + \bar{x}^2)\bar{y} - \bar{x}(s_{xy} + \bar{x}\bar{y}). \tag{1.16}$$

e) Synthesize the previous parts to show that assuming $s_x^2 \neq 0$,

$$\hat{\boldsymbol{x}} = \left(\frac{1}{n}A^T A\right)^{-1}\left(\frac{1}{n}A^T \boldsymbol{y}\right) = \begin{pmatrix} \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} \\ \frac{s_{xy}}{s_x^2} \end{pmatrix}. \tag{1.17}$$

**Problem 1.4** (Pearson correlation coefficient $r$ and the coefficient of determination $R^2$ in simple linear regression)**.** In the previous problem we studied simple linear regression (linear regression with one predictor), where we tried to model the underlying trend in a dataset with $n$ samples $\{(x_i, y_i)\}_{i=1}^n$ by estimating the $i$-th sample via a linear function

$$y_i = \beta_0 + \beta_1 x_i. \tag{1.18}$$

In matrix form, this corresponds to solving the linear system $A\boldsymbol{x} = \boldsymbol{y}$, where

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \ \boldsymbol{x} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{1.19}$$

Since $y \notin \text{Col}(A)$ in general, we can only hope to look for the least squares solution $\hat{\boldsymbol{x}} = (\hat{\beta}_0 \ \hat{\beta}_1)^T$ satisfying $A^T A\hat{\boldsymbol{x}} = A^T\boldsymbol{y}$. We discovered that by assuming $s_x^2 \neq 0$, we have

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{1.20}$$

For linear regression models, one often talks about the coefficient of determination "$R^2$", which measures the predictive power of the model if the model is a good fit.

In this problem we look to explore the geometric interpretation of $R^2$ in the context of simple linear regression. Let $\hat{\boldsymbol{y}} = A\hat{\boldsymbol{x}}$, the projection of $\boldsymbol{y}$ onto the column space $\text{Col}(A)$. The components $\hat{y}_i$ of $\hat{\boldsymbol{y}}$ are often referred to as the "fitted values", as these are the values that lie on the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Define the vector

$$\bar{\boldsymbol{y}} = \bar{y}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix}. \tag{1.21}$$

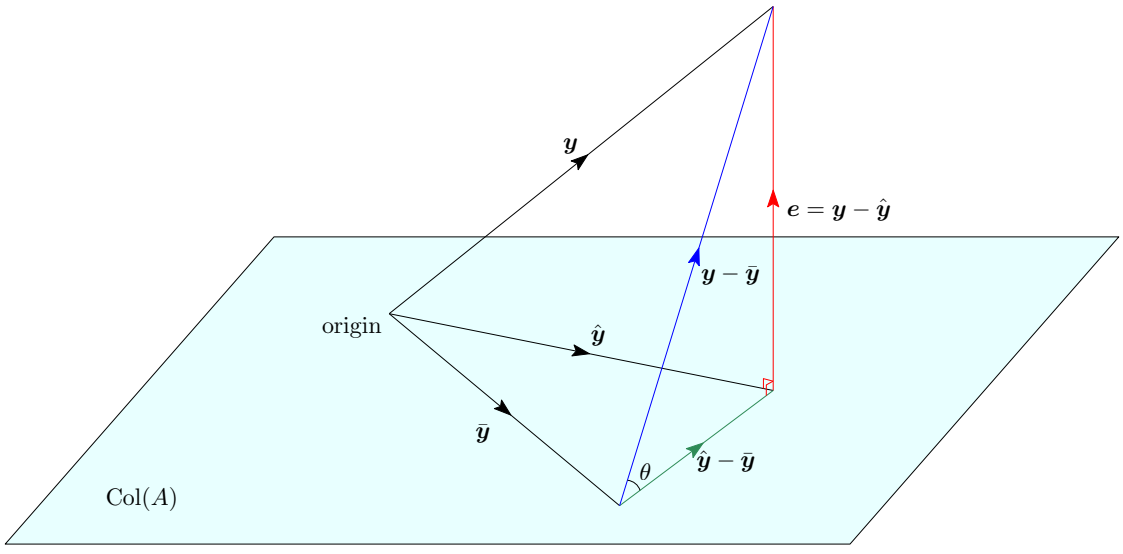
We define the coefficient of determination $R^2$ to be the quantity

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2}. \tag{1.22}$$

We also define the Pearson correlation coefficient $r$ of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ via

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{(\sum_{i=1}^n (y_i - \bar{y})^2)(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2)}} = \frac{s_{y\hat{y}}}{s_y s_{\hat{y}}}, \tag{1.23}$$

where $\frac{n}{n-1} s_{y\hat{y}}$ is the sample covariance between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ defined in the last problem, and $\frac{n}{n-1} s_y^2, \frac{n}{n-1} s_{\hat{y}}^2$ are the sample variances of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$. The goal of this problem is to help you visualize the various quantities defined above using the following figure.



a) Explain why $\boldsymbol{e} \cdot \mathbf{1}_{n\times 1} = 0$. Here $\mathbf{1}_{n\times 1}$ is an $n \times 1$ vector with 1's as its entries, which is also the first column of the matrix $A$. Use this to show that

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}. \tag{1.24}$$

This shows that the average value of the fitted values is equal to the average value of the observed values.

b) Use the previous part to show that

$$r = \frac{(\boldsymbol{y} - \bar{\boldsymbol{y}}) \cdot (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\| \, \|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|} = \cos\theta, \tag{1.25}$$

where $\theta \in [0, \pi]$ is the angle between the vectors $\boldsymbol{y} - \bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}$. This shows that the correlation coefficient between the centered observed values and the centered fitted values has a simple geometric interpretation.

c) Use (1.25) to show that $-1 \leq r \leq 1$.

d) Use the geometry in the figure above to explain why

$$\sin\theta = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|}. \tag{1.26}$$

Use (1.22) and (1.26) to conclude that

$$R^2 = \cos^2\theta = r^2. \tag{1.27}$$

This shows that for simple linear regression, the coefficient of determination $R^2$ is the square of the correlation coefficient $r$.

**Problem 1.5** (Regularization: ridge regression)**.** Suppose we want to study the least squares solution to the linear system $A\boldsymbol{x} = \boldsymbol{b}$, for $A \in \mathcal{M}_{m\times n}(\mathbb{R}), \boldsymbol{b} \in \mathbb{R}^m$. Let $\lambda > 0$, and consider the modified linear system $A_2\boldsymbol{x} = \boldsymbol{b}_2$ for $A_2 \in \mathcal{M}_{(m+n)\times n}(\mathbb{R}), \boldsymbol{b}_2 \in \mathbb{R}^{(m+n)}$ defined via

$$A_2 = \begin{pmatrix} A \\ \sqrt{\lambda} I_{n\times n} \end{pmatrix} = \begin{pmatrix} \boxed{A} \\ \boxed{\begin{matrix} \sqrt{\lambda} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda} \end{matrix}} \end{pmatrix}, \ \boldsymbol{b}_2 = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0}_n \end{pmatrix} = \begin{pmatrix} \boxed{\boldsymbol{b}} \\ \boxed{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}} \end{pmatrix}. \tag{1.28}$$

In other words, the modified system is obtained from augmenting the original linear system with equations of the form $\sqrt{\lambda}x_i = 0$, $1 \le i \le n$. Note that since the unique solution to $(\sqrt{\lambda}I)\boldsymbol{x} = \boldsymbol{0}$ is the zero vector $\boldsymbol{x} = \boldsymbol{0}$, we are in some sense adding in constraints to the original linear system so that large values in the original least squares solution are being "penalized" and being pushed down to 0.

a) Explain why the columns of $A_2$ are always linearly independent, therefore $A_2^T A_2$ is always invertible.

b) Show that

$$\mathbb{R}^n \ni \hat{\boldsymbol{x}}_{\text{ridge}} \text{ solves } A_2^T A_2 \hat{\boldsymbol{x}} = A_2^T \boldsymbol{b}_2 \text{ if and only if } \mathbb{R}^n \ni \hat{\boldsymbol{x}}_{\text{ridge}} \text{ solves } (A^T A + \lambda I)\hat{\boldsymbol{x}} = A^T \boldsymbol{b}. \tag{1.29}$$

c) Show that

$$\|\boldsymbol{b}_2 - A_2\boldsymbol{x}\|^2 = \|\boldsymbol{b} - A\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{x}\|^2, \tag{1.30}$$

therefore the least squares solution $\hat{\boldsymbol{x}}_{\text{ridge}}$ to $A_2\boldsymbol{x} = \boldsymbol{b}_2$ minimizes the modified error function

$$E(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{x}\|^2, \ \boldsymbol{x} \in \mathbb{R}^n. \tag{1.31}$$

In applications, $\hat{\boldsymbol{x}}_{\text{ridge}}$ is referred to as the *ridge regressor* of the linear system $A\boldsymbol{x} = \boldsymbol{b}$, and the interpretation is that it minimizes not the standard least squares error but the modified error defined in (1.31), which is the standard least squares error plus a "regularizing" or "penalty" term. The advantage of performing ridge regression as opposed to standard linear regression is that this method can be applied even if the design matrix has linearly dependent columns.

**Problem 1.6** (Rayleigh quotients and eigenvalues)**.** Suppose $A \in \mathcal{M}_{n\times n}(\mathbb{R})$ is symmetric, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of $A$ and $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ is a corresponding set of orthonormal eigenvectors, which are guaranteed to exist via the spectral theorem. We define the *Rayleigh quotient* $R : \mathcal{M}_{n\times n}(\mathbb{R}) \times (\mathbb{R}^n \setminus \{\mathbf{0}\}) \to \mathbb{R}$ via

$$R(A, \boldsymbol{x}) = \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}. \tag{1.32}$$

Our goal is to show that

$$\lambda_1 = \max_{\boldsymbol{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} R(A, \boldsymbol{x}) \text{ and } \lambda_n = \min_{\boldsymbol{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} R(A, \boldsymbol{x}). \tag{1.33}$$

This gives a *variational characterization* of the eigenvalues of a symmetric matrix in terms of the Rayleigh quotient. Argue as follows.

a) Suppose $\boldsymbol{x} = c_1\boldsymbol{q}_1 + \ldots + c_n\boldsymbol{q}_n$. Show that

$$\boldsymbol{x}^T\boldsymbol{x} = c_1^2 + \ldots + c_n^2 \tag{1.34}$$

and

$$\boldsymbol{x}^T A\boldsymbol{x} = \lambda_1 c_1^2 + \ldots + \lambda_n c_n^2. \tag{1.35}$$

b) Show that

$$(c_1^2 + \ldots + c_n^2)\lambda_n \leq \lambda_1 c_1^2 + \ldots + \lambda_n c_n^2 \leq (c_1^2 + \ldots + c_n^2)\lambda_1. \tag{1.36}$$

c) Use the previous part to show that

$$\lambda_n \leq R(A, \boldsymbol{x}) \leq \lambda_1 \text{ for any } \boldsymbol{x} \neq \mathbf{0}. \tag{1.37}$$

d) Show that if $\boldsymbol{x} = \boldsymbol{q}_1$, then $R(A, \boldsymbol{x}) = \lambda_1$. Also, if $\boldsymbol{x} = \boldsymbol{q}_n$, then $R(A, \boldsymbol{x}) = \lambda_n$.

e) Synthesize the previous parts and show that

$$\lambda_1 = \max_{\boldsymbol{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} R(A, \boldsymbol{x}) \text{ and } \lambda_n = \min_{\boldsymbol{x}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} R(A, \boldsymbol{x}). \tag{1.38}$$

Optional bonus problem: try to extend this idea to characterize singular values of a generic matrix $A \in \mathcal{M}_{m\times n}(\mathbb{R})$.