EECS208 Written HW1

Issued: Sep. 1. **Due:** Sep. 12, 11:59 PM via Gradescope

Reading: Chapters 1, 2, and Appendix A of *High-Dim Data Analysis with Low-Dim Models*.

Problem 1 (ℓ^p-norm)

Given $p \geq 0$, define the function $\|\cdot\|_p : \mathbb{R}^n \mapsto \mathbb{R}$ as

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p},$$

where we slightly abuse the notation by defining $\|x\|_0 = \sum_{i=1}^n |x_i|^0$ and $\|x\|_\infty = \max_{i\in[n]} |x_i|$. Prove that

1. $\forall p \in [0, 1)$, $\|\cdot\|_p$ is *not* a norm of $\mathbb{R}^n$;
2. $\forall p \in \{1, 2, \infty\}$, $\|\cdot\|_p$ is a norm of $\mathbb{R}^n$.

Solutions

1. Prove by contradiction: pick $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^\top, \boldsymbol{e}_2 = [0, 1, \ldots, 0]^\top$ from $\mathbb{R}^n$. Then when $p \in [0, 1)$, we have
$$\|\boldsymbol{e}_1 + \boldsymbol{e}_2\|_p = 2^{1/p} > 2 = \|\boldsymbol{e}_1\|_p + \|\boldsymbol{e}_2\|_p, \tag{0.1}$$
which contradicts the subadditivity.

2. From definition, it is easy to show that $\forall p \in 1, 2, \infty$, $\|\cdot\|_p$ is positive definite and nonnegatively homogeneous, we will only show the subadditivity:
   - When $p = 1$, $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, we have
   $$\|\boldsymbol{a} + \boldsymbol{b}\|_1 = \sum_{i=1}^n |a_i + b_i| \leq \sum_{i=1}^n |a_i| + \sum_{i=1}^n |b_i| = \|\boldsymbol{a}\|_1 + \|\boldsymbol{b}\|_1. \tag{0.2}$$
   - When $p = 2$, $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, we have
   $$\begin{aligned}\|\boldsymbol{a} + \boldsymbol{b}\|_2^2 &= \sum_{i=1}^n (a_i + b_i)^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + 2\sum_{i=1}^n a_i b_i \leq \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + 2\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2 \\ &= (\|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2)^2.\end{aligned} \tag{0.3}$$
   - When $p = \infty$, we have
   $$\|\boldsymbol{a} + \boldsymbol{b}\|_\infty = \max_i |a_i + b_i| = |a_{i^\star} + b_{i^\star}| \leq |a_{i^\star}| + |b_{i^\star}| \leq \max_i |a_i| + \max_i |b_i| = \|\boldsymbol{a}\|_\infty + \|\boldsymbol{b}\|_\infty. \tag{0.4}$$

Problem 2 (Rank-Nullity Theorem)

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, prove the following statements, and suppose bilinear form of the orthogonal complement $\perp$ is defined via Euclidean inner product (Suppose $\mathbb{V} \subseteq \mathbb{R}^n$ is a linear subspace and $\mathbb{V}^\perp$ is the orthogonal complement of $\mathbb{V}$ in $\mathbb{R}^n$, then we have $\langle v, v^\perp \rangle \doteq \sum_{i=1}^n v_i v_i^\perp = 0, \forall v \in \mathbb{V}, v^\perp \in \mathbb{V}^\perp$). Prove that:

1. $\mathsf{null}(\boldsymbol{A})^\perp = \mathsf{range}(\boldsymbol{A}^\top)$
2. $\mathsf{null}(\boldsymbol{A}^\top) = \mathsf{null}(\boldsymbol{A}\boldsymbol{A}^\top)$
3. $\mathsf{dim}(\mathsf{row}(\boldsymbol{A})) + \mathsf{dim}(\mathsf{null}(\boldsymbol{A})) = n$

Solutions

1. $\boldsymbol{x} \in \mathsf{null}(\boldsymbol{A}) \implies \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \implies \boldsymbol{x} \in \mathsf{row}(\boldsymbol{A})^\perp$. Thus we know that $\mathsf{null}(\boldsymbol{A}) = \mathsf{range}(\boldsymbol{A}^\top)^\perp$. Since $S = (S^\perp)^\perp$ holds when S is a subspace of $\mathbb{R}^n$, hence we know that $\mathsf{null}(\boldsymbol{A})^\perp = \mathsf{range}(\boldsymbol{A})$.
2. $\forall \boldsymbol{x} \in \mathsf{null}(\boldsymbol{A}^*)$, $\boldsymbol{A}^*\boldsymbol{x} = \boldsymbol{0} \implies \boldsymbol{A}\boldsymbol{A}^*\boldsymbol{x} = \boldsymbol{0} \implies \boldsymbol{x} \in \mathsf{null}(\boldsymbol{A}\boldsymbol{A}^*)$. On the other hand, $\forall \boldsymbol{x} \in \mathsf{null}(\boldsymbol{A}\boldsymbol{A}^*) \implies \boldsymbol{A}\boldsymbol{A}^*\boldsymbol{x} = \boldsymbol{0} \implies \boldsymbol{x}\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{x} = 0 \implies \|\boldsymbol{A}^*\boldsymbol{x}\|^2 = 0 \implies \boldsymbol{A}^*\boldsymbol{x} = \boldsymbol{0} \implies \boldsymbol{x} \in \mathsf{null}(\boldsymbol{A}^*)$. Thus, we know that $\mathsf{null}(\boldsymbol{A}^*) = \mathsf{null}(\boldsymbol{A}\boldsymbol{A}^*)$.
3. Let us prove a slightly more general version of the last problem: Let $\boldsymbol{A} : X \to Y$ be a linear operator with $\mathsf{dim}(X) = n$. Prove that $\mathsf{dim}(N(\boldsymbol{A})) + \mathsf{dim}(R(\boldsymbol{A})) = n$, i.e., the sum of the dimension of the null space of $\boldsymbol{A}$ and the dimension of the range of $\boldsymbol{A}$ equals the dimension of X.

   Let $\{x_1, \ldots, x_k\} \in X$ be a basis of $N(\boldsymbol{A})$. By basis expansion theorem, we can complete this basis in X as $\{x_1, \ldots, x_k, x_{k+1}, \ldots, x_n\}$. Any vector $x \in X$ can be uniquely represented as

   $$x = \sum_{i=1}^n \alpha_i x_i = \sum_{i=1}^k \alpha_i x_i + \sum_{i=k+1}^n \alpha_i x_i.$$

   After we apply the linear operator $\boldsymbol{A}$ on x, we find out that $\{\boldsymbol{A}(x_{k+1}), ..., \boldsymbol{A}(x_n)\}$ spans the range $R(\boldsymbol{A})$ in Y. Since $x_{k+1}, \ldots, x_n$ are linearly independent from $N(\boldsymbol{A})$, $(\boldsymbol{A}(x_{k+1}), ..., \boldsymbol{A}(x_n))$ should be linearly independent in Y (Please show this yourself by contradiction).

   Hence, $\mathsf{dim}(R(\boldsymbol{A})) = n - k$, which implies $\mathsf{dim}(X) = n = \mathsf{dim}(N(\boldsymbol{A})) + \mathsf{dim}(R(\boldsymbol{A}))$.

Problem 3 (Eigenvalues and Eigenvectors)

Exercise 1.6 of *High-Dim Data Analysis with Low-Dim Models.*

Solutions

According to equation (1.2.20), the first principal component of a random vector $\boldsymbol{y}$ is $\arg\max_{\boldsymbol{u}} \mathsf{Var}(\boldsymbol{u}^\top \boldsymbol{y})$. Notice that

$$\mathsf{Var}(\boldsymbol{u}^\top \boldsymbol{y}) = \mathbb{E}\left[\boldsymbol{u}^\top \boldsymbol{y} - \mathbb{E}(\boldsymbol{u}^\top \boldsymbol{y})\right]^2 = \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]\boldsymbol{u} - (\boldsymbol{u}^\top \boldsymbol{y})^2 = \boldsymbol{u}^\top \boldsymbol{\Sigma}_y \boldsymbol{u}, \tag{0.5}$$

hence,

$$\arg\max_{\boldsymbol{u}} \mathsf{Var}(\boldsymbol{u}^\top \boldsymbol{y}) = \arg\max_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{\Sigma}_y \boldsymbol{u} \tag{0.6}$$

is finding the maximum singular value/vector of $\boldsymbol{\Sigma}_y$.

Proof of Theorem A.29 Since $\boldsymbol{A}$ is symmetric, we can write $\boldsymbol{A}$ as

$$\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top, \tag{0.7}$$

where λ_i are the eigenvalues (actually singular values because $\boldsymbol{A}$ is symmetric). Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Hence, we have

$$\max_{\|\boldsymbol{x}\|_2^2=1} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2^2=1} \boldsymbol{x}^\top \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{x} \leq \lambda_1 \sum_{i=1}^n \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^2 . \tag{0.8}$$

From the definition of SVD, we know that $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n\}$ are an orthonormal basis of $\mathbb{R}^n$. Since $\|\boldsymbol{x}\|_2 = 1$, we have

$$\lambda_1 \sum_{i=1}^n \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle^2 = \lambda_1. \tag{0.9}$$

Hence, we conclude that $\max_{\|\boldsymbol{x}\|_2^2=1} \boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^\top = \lambda_1$. Similarly, we can conclude that $\min_{\|\boldsymbol{x}\|_2^2=1} \boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^\top = \lambda_n$. By repeating the argument in equation (0.8), we can also conclude that λ_k is the optimal value for

$$\max_{\|x\|_2^2=1} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \text{ subject to } x \perp \boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1}. \tag{0.10}$$

Problem 4 (Ridge Regression)

Exercise 1.8 of *High-Dim Data Analysis with Low-Dim Models.*

Solutions

1. Since the objective function is convex, we can consider the critical point of the objective function:

$$\nabla_{\boldsymbol{x}} \left(\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_2^2 \right) = 2\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^\top \boldsymbol{y} + 2\lambda \boldsymbol{x}, \tag{0.11}$$

by setting the gradient to $\mathbf{0}$, yields

$$2\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^\top \boldsymbol{y} + 2\lambda \boldsymbol{x} = \mathbf{0} \implies \boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A}^\top \boldsymbol{y} \tag{0.12}$$

2. $\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}$ is always positive definite for all $\lambda > 0$, because $\forall x \in \mathbb{R}^n$ such that $\boldsymbol{x} \neq \mathbf{0}$, we have

$$\boldsymbol{x}^\top (\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} + \lambda \|\boldsymbol{x}\|_2^2 = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_2^2 > 0, \tag{0.13}$$

since all positive definite matrices are invertible, we know that $\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}$ is invertible.

Problem 5 (Implicit Bias of Gradient Descent)

Exercise 2.10 of *High-Dim Data Analysis with Low-Dim Models.*

Solutions

We can write the iterative formula of the gradient descent as follows:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - 2\alpha \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{y}) = (\boldsymbol{I} - 2\alpha \boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}_k + 2\alpha \boldsymbol{A}^\top \boldsymbol{y}. \tag{0.14}$$

Replacing the above equation with $\boldsymbol{x}_0 = \mathbf{0}$, yields

$$\boldsymbol{x}_k = 2\alpha \left(\sum_{t=0}^{k-1} (\boldsymbol{I} - 2\alpha \boldsymbol{A}^\top \boldsymbol{A})^t \right), \tag{0.15}$$

and when $k \to \infty$, we have

$$\boldsymbol{x}_\infty = 2\alpha \left(\sum_{t=0}^{\infty} (\boldsymbol{I} - 2\alpha \boldsymbol{A}^\top \boldsymbol{A})^t \right), \tag{0.16}$$

Since the matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has full row rank, we can write the matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{U}_1 \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^\top \\ \boldsymbol{V}_2^\top \end{bmatrix}. \tag{0.17}$$

Substituting the SVD into equation (0.16), yields

$$\begin{aligned} \boldsymbol{x}_\infty =& 2\alpha \boldsymbol{V} \left(\sum_{t=0}^{\infty} (\boldsymbol{I} - 2\alpha \boldsymbol{\Sigma}^\top \boldsymbol{\Sigma})^t \right) \boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{y} \\ =& 2\alpha \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix} \left(\sum_{t=0}^{\infty} \begin{bmatrix} \left(\boldsymbol{I} - 2\alpha \boldsymbol{\Sigma}_1^\top \boldsymbol{\Sigma}_1\right)^t & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{\Sigma}_1^\top \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{U}\boldsymbol{y} \\ =& 2\alpha \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix} \left(\sum_{t=0}^{\infty} \begin{bmatrix} \left(\boldsymbol{I} - 2\alpha \boldsymbol{\Sigma}_1^\top \boldsymbol{\Sigma}_1\right)^t \boldsymbol{\Sigma}_1^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right) \boldsymbol{U}\boldsymbol{y}, \end{aligned} \tag{0.18}$$

since $\boldsymbol{\Sigma}_1$ is symmetric and positive definite, we have

$$\boldsymbol{x}_\infty = \boldsymbol{V}_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{U}^\top \boldsymbol{y}. \tag{0.19}$$

Using the Lagrangian dual formulation, the optimal solution to the original optimization problem is

$$\boldsymbol{x}^\star = \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{A}^\top)^{-1} \boldsymbol{y} = \boldsymbol{V}_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{U}\boldsymbol{y}, \tag{0.20}$$

which is exactly the same as $\boldsymbol{x}_\infty$, also since $\boldsymbol{A}$ has full row rank, we have $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_n$.