Math Review (Fall, 2020)
Project

1. (10 points) Consider Linear Regression Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where we observe data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^{p}$. You may view $\boldsymbol{X}$ and $\boldsymbol{y}$ as a constant matrix and vector respectively. We assume $p \leq n$ and $\boldsymbol{X}$ has full rank, that is, $rank(\boldsymbol{X}) = \min\{n, p\} = p$.

(a) (3 points) The ordinary least squares estimator (LSE) $\boldsymbol{\beta}^{(LSE)}$ is the minimizer to ℓ_2 error

$$L_1(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

Find $\boldsymbol{\beta}^{(LSE)}$ and show it is the global minimizer using second derivative test.

(b) (3 points) The ridge regression estimator $\boldsymbol{\beta}^{(ridge)}$ is the minimizer to ℓ_2 error with ℓ_2 regularization

$$L_2(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

where $\lambda > 0$ is a constant. Find the ridge regression estimator $\boldsymbol{\beta}^{(ridge)}$ and show it is the global minimizer using second derivative test.

(c) (4 points) $\boldsymbol{X}\boldsymbol{\beta}^{(OLS)}$ is viewed as a projection of $\boldsymbol{y}$ on the column space of $\boldsymbol{X}$. Let the orthogonal projection matrix be $P_{\boldsymbol{X}}$, write down its formula using $\boldsymbol{X}$ and verify that it is a orthogonal projection matrix, that is, for any $\boldsymbol{u} \in \mathbb{R}^n$, $\langle P_{\boldsymbol{X}}\boldsymbol{u}, \boldsymbol{u} - P_{\boldsymbol{X}}\boldsymbol{u}\rangle = 0$.

2. (10 points) Consider a simple one-layer feedforward neural network model

$$\boldsymbol{y}_i = f_{NN}(\boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i,$$

where $i = 1, \ldots, n$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $\boldsymbol{y}_i \in \mathbb{R}^d$ and

$$f_{NN}(\boldsymbol{x}_i) = B\phi(W\phi(A\boldsymbol{x}_i)).$$

The matrices $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{d \times m}$ and $W \in \mathbb{R}^{m \times m}$. The activation function ϕ is ReLU activation, that is, $\phi(\boldsymbol{u}) = (u_1 \vee 0, u_2 \vee 0, \ldots, u_n \vee 0)^\top$ for $\boldsymbol{u} \in \mathbb{R}^n$. Thus, you may write $D_{i,0}A\boldsymbol{x}_i = \phi(A\boldsymbol{x}_i)$ and $D_{i,1}W\boldsymbol{h}_{i,0} = \phi(W\boldsymbol{h}_{i,0})$ where $D_{i,0}$ and $D_{i,1}$ are diagonal matrices with elements being 0 or 1. Consider the ℓ_2 error

$$L(W) = \frac{1}{2}\sum_{i=1}^{n} \|\boldsymbol{y}_i - f_{NN}(\boldsymbol{x}_i)\|_2^2 .$$

Find the gradient matrix of $L(W)$, $\nabla_W L = \left(\dfrac{\partial L(W)}{\partial W_{s,t}}\right)_{m \times m}$. Be careful with dimensions when you work on vectors and matrices.

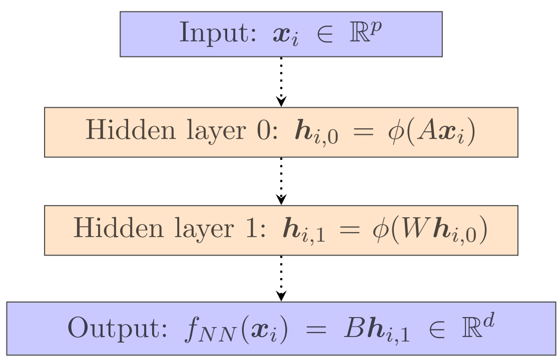

Figure 1: One-layer Neural Network