Part 3 Implicit Regularization / linear regression

Question 1

Prove the following proposition.

Proposition:

With the notations and setting established in class suppose we minimize $L_S(w)$ by initializing $w^{(0)} = a \in \mathbb{R}^d$ and producing iterates $w^{(1)}, w^{(2)}, \ldots$ via iterative algorithm in which every update $w^{(t+1)} - w^{(t)} \in \operatorname{span}\{ \nabla \ell_{(x_i, y_i)}(w) : i \in [m], w \in \mathbb{R}^d \}$.

Assume convergence to global min with zero loss.

Then the sub-optimality of the obtained norm (i.e. the extent to which it is larger than min norm across all global optima) is $\leq \| P_\perp a \|$ where $P_\perp : \mathbb{R}^d \to \mathbb{R}^d$ stands for projection onto the orthogonal complement of $\operatorname{span}\{x_i\}_{i=1}^m$.

Proof

Denote by $\quad w^* = \underset{\substack{w \in \mathbb{R}^d \text{ global min} \\ L_S(w) = 0}}{\arg\min} \|w\|$ the solution

with the min norm across all global optima.

Lemma 1

The solution under the settings in the proposition is

$$\tilde{w} = a + X(X^T X)^{-1}(y - X^T a)$$

With the markings above and as far as lemma 1 is true

we have to prove that

$$\|\tilde{w}\| \leq \|w^*\| + \|P_\perp a\|$$

in order to finish the question.

Proof of lemma 1

First, we'll prove that for every $t \in \mathbb{N}$ $w^t \in \underbrace{a + \text{span}\{x_i\}_{i=1}^m}_{\text{affine subspace}}$

by induction.

base: $w^0 = a = a + \sum_{i=1}^m 0 \cdot x_i \in a + \text{span}\{x_i\}_{i=1}^m$

Step: Suppose $w^t \in a + \text{span}\{x_i\}_{i=1}^m$. We'll prove that $w^{t+1} \in a + \text{span}\{x_i\}_{i=1}^m$

$w^t \in a + \text{span}\{x_i\}_{i=1}^m \Rightarrow w^t = a + \sum_{i=1}^m \alpha_i x_i$ for $\{\alpha_i\}_{i=1}^m$ scalars.

Note that for any $i \in [m]$ and $w \in \mathbb{R}^d$:

$$\nabla \ell (w, x_i, y_i) = (x_i^T w - y_i) \cdot x_i$$

This imlies that for every $t \in \mathbb{N}$

$$w^{t+1} - w^t \in \text{span}\{x_i\}_{i=1}^m$$

(Since we know that $w^{t+1} - w^t \in \text{span}\{\nabla \ell_{(x_i, y_i)}(w) \mid i \in [m]\ w \in \mathbb{R}^d\}$)

$\Rightarrow \quad w^{t+1} - w^t = \sum_{i=1}^m \beta_i x_i$ for some $\{\beta_i\}_{i=1}^m$ scalars.

$\Rightarrow \quad w^{t+1} - (a + \sum_{i=1}^m \alpha_i x_i) = \sum_{i=1}^m \beta_i x_i$

$\Rightarrow \quad w^{t+1} = a + \sum_{i=1}^m (\alpha_i + \beta_i) x_i \quad \Rightarrow w^{t+1} \in a + \text{span}\{x_i\}_{i=1}^m$

$\square$

Hence we get for every $t \in \mathbb{N}$ $w^t \in a + \text{span}\{x_i\}_{i=1}^m$.

Since the subspace $a + \text{span}\{x_i\}_{i=1}^m$ is topologically closed

$\tilde{w} = \lim_{t \to \infty} w^t \in a + \text{span}\{x_i\}_{i=1}^m$

$\Rightarrow \quad \tilde{w} = a + \sum_{i=1}^m r_i x_i$ for some scalars $\{r_i\}_{i=1}^m$

$\Rightarrow \quad \tilde{w} = a + Xr \quad$ for $r = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}$

(continue of the proof next page)

We've got $\tilde{w} = a + Xr$.

Since the convergence is to global min with zero loss we get

$$X^T \tilde{w} = y \Rightarrow X^T(a + Xr) = y$$

$$\Rightarrow X^T a + X^T X r = y$$

$$\Rightarrow X^T X r = y - X^T a$$

$$\text{rank}(X^T X) = m \Rightarrow r = (X^T X)^{-1}(y - X^T a)$$

$$\Rightarrow \tilde{w} = a + X(X^T X)^{-1}(y - X^T a)$$

as required. $\square$

Now, remember that ~~at~~ in class we proved that the solution $w^* = X(X^T X)^{-1} y$ is the one with minimal norm.

Thus we have to show that:

$$\|a + X(X^T X)^{-1}(y - X^T a)\| = \|\tilde{w}\| \leq \|w^*\| + \|P_\perp a\| = \|X(X^T X)^{-1} y\| + \|P_\perp a\|$$

We'll show that:

$$\|a + X(X^T X)^{-1}(y - X^T a)\| = \|a + X(X^T X)^{-1} y - X(X^T X)^{-1} X^T a\|$$

$$\leq \|X(X^T X)^{-1} y\| + \|a - X(X^T X)^{-1} X^T a\|$$

$\triangle$ inequality

$$= \|X(X^T X)^{-1} y\| + \|P_\perp a + P_\parallel a - X(X^T X)^{-1} X^T (P_\perp a + P_\parallel a)\|$$

$a = P_\parallel a + P_\perp a$

where $P_\parallel a \in \text{span}\{x_i\}$ and $P_\perp a$ is the projection as described in the question.

Notice that by definition $\langle x_i, P_\perp a\rangle = 0 \quad \forall i \Rightarrow X^T P_\perp a = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

thus:

$$\textcircled{=} \|X(X^T X)^{-1} y\| + \|P_\perp a + P_\parallel a - X(X^T X)^{-1} X^T P_\parallel a\| \textcircled{=}$$

$P_\parallel a \in \text{span}\{x_i\}_{i=1}^m$, thus $P_\parallel a = X \cdot b$ for some $b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$

$$\textcircled{=} \|X(X^T X)^{-1} y\| + \|P_\perp a + Xb - X\underline{(X^T X)^{-1}}\,\underline{X^T X} b\|$$

$$= \|X(X^T X)^{-1} y\| + \|P_\perp a\|$$

Which means that the suboptimality of the obtained norm $\leq \|P_\perp a\|$

as required. $\square$