Discussion 9

1. Gradient Descent for Matrices of Full Row Rank

Consider a matrix $X \in \mathbb{R}^{n \times d}$ with $n < d$ and a vector $\vec{y} \in \mathbb{R}^n$, both of which are known and given to you. Suppose X has full row rank.

(a) Consider the following problem:

$$X\vec{w} = \vec{y} \tag{1}$$

where $\vec{w} \in \mathbb{R}^d$ is unknown. How many solutions does (1) have? *Justify your answer.*

(b) Consider the minimum-norm problem

$$\vec{w}_\star = \operatorname*{argmin}_{\substack{\vec{w} \in \mathbb{R}^d \\ X\vec{w} = \vec{y}}} \|\vec{w}\|_2^2 . \tag{2}$$

We know that the optimal solution to this problem is $\vec{w}_\star = X^\top (XX^\top)^{-1}\vec{y}$. Now let $X = U\Sigma V^\top = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$ be the SVD of X, where $\Sigma_1 \in \mathbb{R}^{n \times n}$. Recall that this is possible because $n < d$ and X is full row rank. Prove that $\vec{w}_\star$ is given by

$$\vec{w}_\star = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{3}$$

(c) Let $\eta > 0$, and I be the identity matrix of the appropriate dimension. Using the SVD $X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$, prove the following identity for all positive integers $i > 0$:

$$(I - \eta X^\top X)^i = V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top. \tag{4}$$

(d) Recall that $X \in \mathbb{R}^{n \times d}$, and that we can write the SVD of X as $X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$. We will use gradient descent to solve the minimization problem

$$\min_{\vec{w} \in \mathbb{R}^d} \frac{1}{2} \|X\vec{w} - \vec{y}\|_2^2, \tag{5}$$

with step-size $\eta > 0$. Let $\vec{w}_0 = \vec{0}$ be the initial state, and $\vec{w}_k$ be the k^{th} iterate of gradient descent. Use the identity:

$$(I - \eta X^\top X)^i = V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top. \tag{6}$$

to prove that after k steps, we have

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{7}$$

HINT: Remember to set $\vec{w}_0 = \vec{0}$.

(e) Now let $0 < \eta < \frac{1}{\sigma_1^2}$, where σ_1 denotes the maximum singular value of $X = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top$. Let $\vec{w}_k$ be given as

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{8}$$

and let $\vec{w}_\star$ be the minimum norm solution given as

$$\vec{w}_\star = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{9}$$

Prove that $\lim_{k\to\infty} \vec{w}_k = \vec{w}_\star$.

HINT: You may use the following result without proof. When all eigenvalues of $A \in \mathbb{R}^{n\times n}$ have magnitude < 1, we have the identity $(I - A)^{-1} = I + A + A^2 + \ldots$.

2. Stochastic Gradient Method

Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, with domain $\mathbb{R}^n$, whose minimum we seek to find, we could use the gradient descent algorithm $\vec{\theta}_{k+1} = \vec{\theta}_k - \eta \nabla f(\vec{\theta}_k)$, with fixed step size $\eta > 0$, starting from an initial condition $\vec{\theta}_0 \in \mathbb{R}^n$. As we have seen, there is no guarantee that this algorithm converges, and even if it does it may only converge to a local minimum of the function.

One issue with the gradient descent algorithm is the complexity of computing the gradient at each time step. If the function could be decomposed as a summation of multiple functions $f(\vec{\theta}) = \sum_{l=1}^{m} f_l(\vec{\theta})$, for each of which the gradient is easily computable, then we can use the *stochastic gradient* method. For instance, the squared-error-loss function which shows up in the least squares problem is well-suited for minimization with the stochastic gradient method. Here our problem is

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2}\|X\vec{\theta} - \vec{y}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(\vec{x}_i^\top \vec{\theta} - y_i)^2,$$

where $\vec{x}_i^T$ is the i-th row of $X \in \mathbb{R}^{m \times n}$, and $\vec{y} \in \mathbb{R}^m$ (recall that the rows of X are the transposes of the *feature vectors* and the entries of $\vec{y}$ are the corresponding *responses*). We can write this objective function as $f(\vec{\theta}) = \sum_{i=1}^{m} f_i(\vec{\theta})$, with

$$f_i(\vec{\theta}) := \frac{1}{2}(\vec{x}_i^\top \vec{\theta} - y_i)^2, \quad \text{for } i = 1, \ldots, m.$$

Then the stochastic gradient method gives the update rule

$$\vec{\theta}_{k+1} = \vec{\theta}_k - \eta_k \nabla f_{s[k]}(\vec{\theta}_k),$$

where η_k is the step size at time $k \in \mathbb{N}$, and $s[k] \in \{1, \ldots, m\}$ is the index of the component function chosen at time k in order to decide the update. The value of $s[k]$ is usually chosen by drawing a number at random from the set $\{1, \ldots, m\}$, or by randomly shuffling this set and going over it sequentially in cyclic order. However this choice is done, we will assume that each $i \in \{1, \cdots, n\}$ is chosen infinitely often.

(a) Assume that $\{\vec{x}_i\}_{i=1}^m$ is a set of mutually orthogonal vectors. Find a fixed step size η so that the stochastic gradient method converges to a solution of the least squares problem.

(b) If we no longer assume $\{\vec{x}_i\}_{i=1}^m$ is orthogonal, can we still find a fixed step size small enough that the stochastic gradient method converges?