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.*

Solution: Since $\vec{y}$ is in the range of X, this implies that there exists $\vec{w}_0$ such that $\vec{y} = X\vec{w}_0$. Now let $\vec{s}$ be any non-zero vector in the null space of X (which exists since $\dim(\mathcal{N}(X)) = d - n > 0$), and consider an arbitrary vector $\vec{w}_{\text{new}} = \vec{w}_0 + t\vec{s}$, where $t \in \mathbb{R}$. Since $X\vec{w}_{\text{new}} = X\vec{w}_0 = \vec{y}$, we conclude that there are infinitely many solutions, corresponding to infinitely many values of t. In fact, every solution $\vec{w}$ can be written as $\vec{w}_0 + \vec{z}$, for some $\vec{z} \in \mathcal{N}(X)$.

(b) Consider the minimum-norm problem

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

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

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

Solution: By plugging in the SVD of X in the expression of $\vec{w}_\star$, we have

$$\begin{aligned}
\vec{w}_\star &= X^\top (XX^\top)^{-1}\vec{y} && (4)\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \left(U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} V^\top V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \right)^{-1} \vec{y}, && \text{(plugged in the SVD of } X\text{)}\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \left(U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \right)^{-1} \vec{y}, && \text{(by } V^\top V = I\text{)}\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top U \left(\begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \right)^{-1} U^\top \vec{y}, && \text{(by } U^{-1} = U^\top\text{)}\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \left(\begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \right)^{-1} U^\top \vec{y}, && \text{(by } U^\top U = I\text{)}\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \left(\Sigma_1^2 \right)^{-1} U^\top \vec{y}, && \text{(took the matrix product of } \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix}\text{)}\\
&= V \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \Sigma_1^{-2} U^\top \vec{y}, && \text{(}\Sigma_1 \text{ is a square matrix and invertible)}\\
&= V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. && (5)
\end{aligned}$$

(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 $i > 0$:

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

Solution: We have

$$\begin{aligned}
(I - \eta X^\top X)^i &= \left(I - \eta (U \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} V^\top)^\top (U \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} V^\top)\right)^i, && \text{(plugged in the SVD of } X\text{)} \\
&= \left(I - \eta V \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix} U^\top U \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} V^\top\right)^i, && \text{(took the transpose of } U \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} V^\top\text{)} \\
&= \left(I - \eta V \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix} \begin{bmatrix}\Sigma_1 & 0\end{bmatrix} V^\top\right)^i, && \text{(by } U^\top U = I\text{)} \\
&= \left(I - \eta V \begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix} V^\top\right)^i, && \text{(took the matrix product of } \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix} \begin{bmatrix}\Sigma_1 & 0\end{bmatrix}\text{)} \\
&= \left(VV^\top - \eta V \begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix} V^\top\right)^i, && \text{(by } I = VV^\top\text{)} \\
&= \left(V \left(I - \eta \begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix}\right) V^\top\right)^i, && \text{(combine the diagonal matrices)} \\
&= V \left(I - \eta \begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix}\right)^i V^\top. && \text{(by applying } V^\top V = I \text{ repeatedly)}
\end{aligned}$$

(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

$$\min_{\vec{w} \in \mathbb{R}^d} \frac{1}{2} \|X\vec{w} - \vec{y}\|_2^2, \tag{7}$$

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:

$$(I - \eta X^\top X)^i = V \left(I - \eta \begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix}\right)^i V^\top. \tag{8}$$

to prove that after k steps, we have

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix}\Sigma_1^2 & 0 \\ 0 & 0\end{bmatrix}\right)^i \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix} U^\top \vec{y}. \tag{9}$$

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

Solution: With $\nabla_{\vec{w}} f(\vec{w}) = X^\top (X\vec{w} - \vec{y})$, the gradient updates are of the form:

$$\vec{w}_{k+1} = \vec{w}_k - \eta \nabla_{\vec{w}} f(\vec{w}_k) \tag{10}$$

$$= (I - \eta X^\top X)\vec{w}_k + \eta X^\top \vec{y} \tag{11}$$

$$\implies \vec{w}_k = (I - \eta X^\top X)^k \vec{w}_0 + \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y} \tag{12}$$

$$= \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y}. \tag{13}$$

Using the identity given, we have

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} (I - \eta X^\top X)^i X^\top \vec{y} \tag{14}$$

$$= \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i V^\top \left(V \Sigma^\top U^\top \right) \vec{y} \tag{15}$$

$$= \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \Sigma^\top U^\top \vec{y} \tag{16}$$

$$= \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{17}$$

(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

$$\vec{w}_k = \eta \sum_{i=0}^{k-1} V \left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \right)^i \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{18}$$

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

$$\vec{w}_\star = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}. \tag{19}$$

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 $(I - A)^{-1} = I + A + A^2 + \ldots$.

Solution: We start with (9) and simplify, obtaining

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

Taking limits, we have

$$\begin{aligned}
\lim_{k\to\infty} \vec{w}_k &= \eta V \begin{bmatrix} \sum_{i=0}^{\infty} (I - \eta\Sigma_1^2)^i \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y} \\
&= \eta V \begin{bmatrix} (I - (I - \eta\Sigma_1^2))^{-1} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}, && \text{(applied the identity in the hint on } I - \eta\Sigma_1^2\text{)} \\
&= \eta V \begin{bmatrix} (\eta\Sigma_1^2)^{-1} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y}, && \text{(}\Sigma_1^2 \text{ is a square matrix and invertible)} \\
&= \eta V \begin{bmatrix} \frac{1}{\eta}\Sigma_1^{-2} \Sigma_1 \\ 0 \end{bmatrix} U^\top \vec{y} \\
&= V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^\top \vec{y}
\end{aligned}$$

as desired. Here the infinite sum is evaluated as in the hint because the eigenvalues of $I - \eta\Sigma_1^2$ are all in the interval $(0, 1) \subseteq (-1, 1)$. Indeed, the eigenvalues of $I - \eta\Sigma_1^2$ are $1 - \eta\sigma_i^2$, where σ_i are the entries of Σ_1 and thus the nonzero singular values of X. Since $\sigma_i > 0$, we know $1 - \eta\sigma_i^2 < 1$. Now, since $\eta < \frac{1}{\sigma_1^2}$, we have $1 - \eta\sigma_i^2 > 1 - \frac{\sigma_i^2}{\sigma_1^2} \geq 0$. Thus the eigenvalues of $I - \eta\Sigma_1^2$ are contained in $(-1, 1)$ and the hint applies.

A common error, is to apply the hint directly on $\left(I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix}\right)$. Note that the eigenvalues of

$$I - \eta \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I - \eta\Sigma_1^2 & 0 \\ 0 & I \end{bmatrix}$$

are in the interval $(0, 1]$, which breaks the condition we made on the A matrix described in the hint, all eigenvalues of A having magnitude strictly < 1.

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 $i \in \{1, \cdots, m\}$ 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.

Solution: For SGD, if $\vec{x}_i$ is the point drawn at iteration $k \in \mathbb{N}$, we can write the update rule as

$$\begin{aligned}\vec{\theta}_{k+1} &= \vec{\theta}_k - \frac{1}{2}\eta \nabla_{\vec{\theta}_k}(\vec{x}_i^\top \vec{\theta}_k - y_i)^2 \\ &= \vec{\theta}_k - \eta(\vec{x}_i^\top \vec{\theta}_k - y_i)\vec{x}_i.\end{aligned}$$

To analyze the convergence, we consider the dynamics of the error terms $e_k^{(i)} := \vec{x}_i^T \vec{\theta}_k - y_i$, where $e_k^{(i)}$ denotes the error associated with data point i at time k, $1 \le i \le m$. For any $1 \le j \le m$, we have

$$\vec{\theta}_{k+1}^T \vec{x}_j - y_j = \vec{\theta}_k^T \vec{x}_j - y_j - \eta(\vec{x}_i^\top \vec{\theta}_k - y_i)\vec{x}_i^T \vec{x}_j,$$

which shows that (because $\{\vec{x}_i\}_{i=1}^m$ is a mutually-orthogonal set of vectors) we have

$$e_{k+1}^{(i)} = \begin{cases} e_k^{(i)} & \text{if } \vec{x}_i \text{ is not drawn at time } k, \\ (1 - \eta\|\vec{x}_i\|_2^2)e_k^{(i)} & \text{if } \vec{x}_i \text{ is drawn at time } k. \end{cases}$$

If $0 < \eta < 2/\max_i \|\vec{x}_i\|_2^2$ then we have $-1 < 1 - \eta\|\vec{x}_i\|_2^2 < 1$ for all i. Hence, if every point in the set $\{\vec{x}_i\}_{i=1}^m$ is drawn infinitely often, all the error terms go to zero, so the objective function converges to zero, which is its optimal value.

To show that $\vec{\theta}_k$ also converges, we also consider the dynamics of $\vec{\theta}_k$. Assume the initialization of θ is

$$\vec{\theta}_0 = \sum_{i=1}^{m} \alpha_i \vec{x}_i + \sum_{i=1}^{n-m} \beta_i \vec{z}_i,$$

where $\{\vec{z}_i\}_{i=1}^{n-m}$ is a set of vectors orthogonal to $\{\vec{x}_i\}_{i=1}^{m}$ which span $\mathbb{R}^n$ along with $\{\vec{x}_i\}_{i=1}^{m}$. Then the update rule for $\vec{\theta}_k$ shows that α_i converges to its optimal value for all $i = 1, \ldots, m$, whereas β_i remains at their initial value for all $i = 1, \ldots, n-m$. Therefore, $\vec{\theta}_k$ also converges (to a point that in general depends on the initial point $\vec{\theta}_0$).

(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?

Solution: Assume that $\{\vec{x}_i\}_{i=1}^m$ is not an orthogonal set of vectors. If $\vec{x}_i$ is drawn at time k, the error term for point j becomes:

$$\vec{\theta}_{k+1}^\top \vec{x}_j - y_j = \vec{\theta}_k^\top \vec{x}_j - y_j - \eta(\vec{x}_i^\top \vec{\theta}_k - y_i)\vec{x}_i^\top \vec{x}_j,$$

which we can write as

$$e_{k+1}^{(j)} = e_k^{(j)} - \eta e_k^{(i)} \vec{x}_i^\top \vec{x}_j. \tag{20}$$

Suppose $\vec{x}_i^\top \vec{x}_j \neq 0$. Then if $\{e_k^{(j)}, k \geq 0\}$ converges as $k \to \infty$, (20) tells us that $e_k^{(i)}$ must converge to 0 as $k \to \infty$. Similarly, by interchanging the roles of i and j, we can conclude that $e_k^{(j)}$ must converge to 0 as $k \to \infty$.

Thus, we have to have

$$e_k^{(i)}, e_k^{(j)} \to 0 \quad \text{for every } i, j \text{ such that } \vec{x}_i^\top \vec{x}_j \neq 0.$$

However, it is possible that the error terms $e_k^{(i)}$, $e_k^{(j)}$ cannot be both zero due to inconsistency in y_i and y_j; for example, when $\vec{x}_j = 2\vec{x}_i$ but $y_j \neq 2y_i$. Therefore it may happen that the stochastic gradient method with fixed step size does not converge, no matter how small the step size is chosen. In general, diminishing the step size over time is required to ensure convergence.