CS 188
Summer 2023

Discussion 6C Solutions

1 Vector Calculus

Let $\vec{x}, \vec{c} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. For the following parts, before taking any derivatives, identify what the derivative looks like (is it a scalar, vector, or matrix?) and how we calculate each term in the derivative. Then carefully solve for an arbitrary entry of the derivative, then stack/arrange all of them to get the final result. Note that the convention we will use going forward is that vector derivatives of a scalar (with respect to a column vector) are expressed as a row vector, i.e. $\frac{\partial f}{\partial \vec{x}} = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}]$ since a row acting on a column gives a scalar. You may have seen alternative conventions before, but the important thing is that you need to understand the types of objects and how they map to the shapes of the multidimensional arrays we use to represent those types.

1. Show $\frac{\partial}{\partial \vec{x}}(\vec{x}^T \vec{c}) = \vec{c}^T$ This is a vector derivative of a scalar quantity, so our result will be a row vector. Looking at the i-th entry, $\frac{\partial}{\partial x_i}(\vec{x}^T \vec{c}) = \frac{\partial}{\partial x_i}(\sum_j c_j x_j) = c_i$. Stacking all the entries into a row vector, we get $\vec{c}^T$.

2. Show $\frac{\partial}{\partial \vec{x}}||\vec{x}||_2^2 = 2\vec{x}^T$ This is a vector derivative of a scalar quantity, so our result will be a row vector. Looking at the i-th entry, $\frac{\partial}{\partial x_i}(||\vec{x}||_2^2) = \frac{\partial}{\partial x_i}(\sum_j x_j^2) = 2x_i$. Stack all the entries into a row to get $2\vec{x}^T$.

3. Show $\frac{\partial}{\partial \vec{x}}(A\vec{x}) = A$ This is a vector derivative of a vector quantity, so the result will be a matrix. Let $\vec{f} = A\vec{x}$. Note that $f_i = \sum_k A_{ik} x_k$ Looking at the (i, j)-th entry of our matrix, $\frac{\partial f_i}{\partial x_j} = \frac{\partial}{\partial x_j}(\sum_k A_{ik} x_k) = A_{ij}$. Arranging all of these in a matrix will recover A.

4. Show $\frac{\partial}{\partial \vec{x}}(\vec{x}^T A \vec{x}) = \vec{x}^T(A + A^T)$ This is a vector derivative of a scalar quantity, so our result will be a vector. Before taking any derivatives, we can write $\vec{x}^T A \vec{x} = \sum_i \sum_j A_{ij} x_i x_j$. Taking the derivative with respect to an arbitrary x_k and focusing on just the terms involving x_k (as the derivative of the other terms wrt x_k is zero), we can write

$$\frac{\partial}{\partial x_k}(\vec{x}^T A \vec{x}) = \frac{\partial}{\partial x_k}((\sum_{j \neq k} A_{kj} x_k x_j) + (\sum_{i \neq k} A_{ik} x_i x_k) + A_{kk} x_k^2)$$

$$= (\sum_{j \neq k} A_{kj} x_j) + (\sum_{i \neq k} A_{ik} x_i) + 2A_{kk} x_k$$

$$= (\sum_j A_{kj} x_j) + (\sum_i A_{ik} x_i) = \sum_i (A_{ki} x_i + A_{ik} x_i)$$

$$= \vec{x}^T(\text{kth row of A} + \text{kth col of A}) = \vec{x}^T(A_k + A_k^T)$$

Stacking all the results in a row vector, we get $\frac{\partial}{\partial \vec{x}}(\vec{x}^T A \vec{x}) = \vec{x}^T(A + A^T)$ as desired.

5. Under what condition is the previous derivative equal to $2\vec{x}^T A$? We want $(A + A^T) = 2A$. This is true if and only if $A = A^T$, ie. the matrix A is symmetric.

2 Solving Linear Regression with Vector Calculus

In this problem we will solve two variations of linear regression – ordinary least squares and ridge regression – using vector calculus.

1. *Ordinary Least Squares* Consider the equation $X\vec{w} = \vec{y}$, where $X \in \mathbb{R}^{n \times d}$ is a non-square data matrix, $w \in \mathbb{R}^d$ is a weight vector, and $y \in \mathbb{R}^n$ is vector of labels corresponding to the datapoints in each row of X.

   Consider the case where $n > d$, i.e. our data matrix X has more rows than columns (tall matrix) and the system is overdetermined. **How do we find the weights $\vec{w}$ that minimizes the error between $X\vec{w}$ and y?** In other words, we want to solve $\min_{\vec{w}} \|X\vec{w} - \vec{y}\|^2$.

   Use vector calculus to solve this optimization problem for $\vec{w}$.

   Call our objective f. Expand

   $$f(\vec{w}) = \vec{w}^T X^T X \vec{w} - 2\vec{w}^T X\vec{y} + \vec{y}^T\vec{y}$$

   Take gradient wrt $\vec{w}$ and set it to zero:

   $$\nabla_{\vec{w}} f = 2X^T X\vec{w} - 2X^T\vec{y} = \vec{0}$$

   Solve for $\vec{w}$:

   $$(X^T X)\vec{w} = X^T\vec{y}$$
   $$\vec{w} = (X^T X)^{-1} X^T\vec{y}$$

2. *Ridge Regression* Ridge regression can be understood as the unconstrained optimization problem

   $$\arg\min_{\vec{w}} \|\vec{y} - X\vec{w}\|_2^2 + \lambda\|\vec{w}\|_2^2, \tag{1}$$

   where $X \in \mathbb{R}^{n \times d}$ is a data matrix, and $\vec{y} \in \mathbb{R}^n$ is the target vector of measurement values. What's new compared to the simple OLS problem is the addition of the $\lambda\|\vec{w}\|^2$ term, which can be interpreted as a "penalty" on the weights being too big.

   Use vector calculus to expand the objective and solve this optimization problem for $\vec{w}$.

   Call our objective f. Expand

   $$f(\vec{w}) = \vec{w}^T X^T X\vec{w} - 2\vec{w}^T X\vec{y} + \vec{y}^T\vec{y} + \lambda\vec{w}^T\vec{w}$$

   Take gradient wrt $\vec{w}$ and set it to zero:

   $$\nabla_{\vec{w}} f = 2X^T X\vec{w} - 2X^T\vec{y} + 2\lambda\vec{w} = \vec{0}$$

   Solve for $\vec{w}$:

   $$(X^T X + \lambda I)\vec{w} = X^T\vec{y}$$
   $$\vec{w} = (X^T X + \lambda I)^{-1} X^T\vec{y}$$