Discussion Week 3

September 13, 2024

1 Matrix Calculus

1.1 The Gradient

Let $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ be a function that takes a matrix A of size $m \times n$ and returns a real value. The gradient of f with respect to $A \in \mathbb{R}^{m \times n}$ is defined as the matrix of partial derivatives:

$$\nabla_A f(A) \in \mathbb{R}^{m \times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}.$$

If A is a vector $x \in \mathbb{R}^n$, then:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix}^\top.$$

1.2 The Hessian

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real value. The Hessian matrix, denoted as $\nabla_x^2 f(x)$, is the matrix of second-order partial derivatives:

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

The Hessian is always symmetric, i.e., $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$.

1.3 Gradients and Hessians of Quadratic Forms

For a quadratic function $f(x) = x^\top Ax$ where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, the gradient and Hessian are:

$$\nabla_x f(x) = 2Ax, \quad \text{and} \quad \nabla_x^2 f(x) = 2A.$$

This can be derived as follows:

$$\nabla_x (x^\top Ax) = \nabla_x \left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \right) = 2Ax.$$

1.4 Matrix Calculus for Determinants

For a square matrix $A \in \mathbb{R}^{n\times n}$, the gradient of its determinant with respect to A is:

$$\nabla_A |A| = |A| A^{-\top}.$$

Similarly, for the function $f(A) = \log |A|$ (where A is a positive definite matrix), the gradient is given by:

$$\nabla_A \log |A| = A^{-1}.$$

1.5 Gradients and Hessians of Quadratic and Linear Functions

Now let's try to determine the gradient and Hessian matrices for a few simple functions. It should be noted that all the gradients given here are special cases of the gradients given in the CS229 lecture notes.

1.6 Linear Function Gradient

For $x \in \mathbb{R}^n$, let $f(x) = b^\top x$ for some known vector $b \in \mathbb{R}^n$. Then:

$$f(x) = \sum_{i=1}^{n} b_i x_i.$$

Taking the partial derivative with respect to x_k:

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k.$$

Thus, the gradient of $f(x)$ is:

$$\nabla_x b^\top x = b.$$

This should be compared to the analogous situation in single variable calculus, where $\partial/\partial x \, (ax) = a$.

1.7 Quadratic Function Gradient

Now consider the quadratic function $f(x) = x^\top Ax$ for a symmetric matrix $A \in \mathbb{R}^{n \times n}$. Recall that:

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

Taking the partial derivative of $f(x)$ with respect to x_k, we have:

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

Let's separate out the terms involving x_k and x_k^2 separately:

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right].$$

Calculating each term:

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

Simplifying by combining symmetric terms (since A is symmetric, $A_{ik} = A_{ki}$):

$$\frac{\partial f(x)}{\partial x_k} = \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j = 2 \sum_{i=1}^{n} A_{ki} x_i = 2(Ax)_k.$$

Thus, the gradient of $f(x) = x^\top Ax$ is:

$$\nabla_x f(x) = 2Ax.$$

This result is analogous to the single-variable calculus result $\partial / \partial x \, (ax^2) = 2ax$.

1.8 Quadratic Function Hessian

Finally, let's look at the Hessian of the quadratic function $f(x) = x^\top Ax$. It should be obvious that the Hessian of a linear function $b^\top x$ is zero. In this case:

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_\ell} \left[\frac{\partial f(x)}{\partial x_k} \right] = \frac{\partial}{\partial x_\ell} \left[2 \sum_{i=1}^{n} A_{ki} x_i \right] = 2A_{k\ell}.$$

Therefore, the Hessian is:

$$\nabla_x^2 f(x) = 2A.$$

This should be entirely expected and is analogous to the single-variable fact that $\partial^2/\partial x^2\,(ax^2) = 2a$.

1.9 Summary

To recap:

- $\nabla_x b^\top x = b$.
- $\nabla_x x^\top A x = 2Ax$ (if A is symmetric).
- $\nabla_x^2 x^\top A x = 2A$ (if A is symmetric).

1.10 Least Squares

Let's apply the equations we obtained in the last section to derive the least squares equations. Suppose we are given matrices $A \in \mathbb{R}^{m\times n}$ (for simplicity, we assume A is full rank) and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$ (the range space of A). In this situation, we will not be able to find a vector $x \in \mathbb{R}^n$ such that $Ax = b$. Instead, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $\|Ax - b\|_2^2$.

Using the fact that $\|x\|_2^2 = x^T x$, we have

$$\|Ax - b\|_2^2 = (Ax - b)^\top (Ax - b).$$

Expanding this, we get:

$$\|Ax - b\|_2^2 = x^\top A^\top Ax - 2b^\top Ax + b^\top b.$$

Taking the gradient of $x^\top A^\top Ax - 2b^\top Ax + b^\top b$ with respect to x, and using the properties derived in the previous section, we have:

$$\nabla_x \left(x^\top A^\top Ax - 2b^\top Ax + b^\top b\right) = \nabla_x \left(x^\top A^\top Ax\right) - \nabla_x \left(2b^\top Ax\right) + \nabla_x \left(b^\top b\right).$$

Calculating each term separately:
- **Gradient of the first term**:

$$\nabla_x \left(x^\top A^\top Ax\right) = 2A^\top Ax.$$

- **Gradient of the second term**:

$$\nabla_x \left(-2b^\top Ax\right) = -2A^\top b.$$

- **Gradient of the third term:** Since $b^\top b$ is a constant with respect to x, its gradient is zero:

$$\nabla_x \left(b^\top b\right) = 0.$$

Setting the gradient equal to zero and solving for x:

$$2A^\top Ax - 2A^\top b = 0.$$

Simplifying:

$$A^\top Ax = A^\top b.$$

This is known as the **normal equation**. Solving for x:

$$x = (A^\top A)^{-1} A^\top b,$$

which is the same expression we derived in class.

2 References

1. Zico Kolter, Chuong Do, *CS229 Linear Algebra Review and Reference*, Stanford University, 2012.