Gradient and Hessian

Konstantin Tretyakov

1 Gradient

Definition Let $f : \mathbb{R}^n \to \mathbb{R}^m$. We say, that f is differentiable at point $\mathbf{x}_0$, if there exists a linear transformation $\mathbf{A}(\mathbf{x}_0)$, such that

$$f(\mathbf{x}_0 + \mathbf{\Delta x}) = f(\mathbf{x}_0) + \mathbf{A}(\mathbf{x}_0)\mathbf{\Delta x} + o(\mathbf{\Delta x})$$

We call a function f differentiable on a set $Q \subset \mathbb{R}^n$, if it is differentiable at each point of Q. If f is differentiable on $\mathbb{R}^n$, we just say that f is differentiable. The matrix $\mathbf{A}(\mathbf{x}_0)$ is referred to as the *derivative* or *Jacobi matrix* of f (at point $\mathbf{x}_0$), and is denoted by $\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{x}}$ or $f'(\mathbf{x}_0)$.

Exercise: What does $o(\mathbf{\Delta x})$ mean in the previous definition and why isn't it $o(\|\mathbf{\Delta x}\|)$?

Theorem 1.1

$$(\mathbf{A}(\mathbf{x}_0))_{ij} = \frac{\partial f_i(\mathbf{x}_0)}{\partial x_j}$$

Exercise*: Prove it.
Hint: First consider $f : \mathbb{R}^n \to \mathbb{R}$ and use the total differential formula $\mathrm{d}f(\mathbf{x}) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}\mathrm{d}x_i$

Theorem 1.2

$$(f(g(\mathbf{x}))' = f'(g(\mathbf{x}))g'(\mathbf{x})$$

Exercise*: Prove it.
Hint: First consider $f : \mathbb{R}^n \to \mathbb{R}$ and use the chain rule $\frac{\partial f \circ g}{\partial x_j}(\mathbf{x}) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(g(\mathbf{x})) \cdot \frac{\partial g_i}{\partial x_j}(\mathbf{x})$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Its derivative (at point $\mathbf{x}_0$) is then a $1 \times n$ matrix

$$\mathbf{A}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f(\mathbf{x}_0)}{\partial x_1} & \frac{\partial f(\mathbf{x}_0)}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x}_0)}{\partial x_n} \end{pmatrix}$$

The vector $\mathbf{A}(\mathbf{x}_0)^T$ is referred to as the *gradient* of f and is denoted as $\mathrm{grad}_{\mathbf{x}} f(\mathbf{x}_0)$ or $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$ or simply $\nabla f(\mathbf{x}_0)$.

Theorem 1.3

$$\begin{aligned} \nabla(f(\mathbf{x})g(\mathbf{x})) &= \nabla f(\mathbf{x})g(\mathbf{x}) + f(\mathbf{x})\nabla g(\mathbf{x}) \\ \nabla(f(\mathbf{x})/g(\mathbf{x})) &= (\nabla f(\mathbf{x})g(\mathbf{x}) - f(\mathbf{x})\nabla g(\mathbf{x}))/g^2(\mathbf{x}) \\ \nabla(f(g(\mathbf{x}))) &= \nabla f(g(\mathbf{x}))\nabla g(\mathbf{x}) = f'(g(\mathbf{x}))\nabla g(\mathbf{x}) \end{aligned}$$

Exercise: Prove it.
Hint: Use the theorems (1.1) and (1.2).

Exercise*: Convince yourself, that gradient can be regarded as a vector pointing to the direction of the steepest ascent on the surface of the function.

Examples

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^T$$
$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A}+\mathbf{A}^T)$$

Exercise: Prove these.

2 Hessian

Definition Let $f:\mathbb{R}^n \to \mathbb{R}$ be differentiable and $\nabla f:\mathbb{R}^n \to \mathbb{R}^n$ differentiable (i.e. let f be twice differentiable). The derivative of ∇f at point $\mathbf{x}_0$ is an $n \times n$ matrix $\mathbf{H}(\mathbf{x}_0)$, which is referred to as the *second derivative* or *Hessian* of f (at point $\mathbf{x}_0$) and is denoted as $\frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 \mathbf{x}}$ or $\nabla^2 f(\mathbf{x}_0)$.

Theorem 2.1
$$(\mathbf{H}(\mathbf{x}_0))_{ij} = \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j}$$

Exercise: Prove it.
Hint: Statement follows from (1.1).

Theorem 2.2 *If partial derivatives* $\frac{\partial^2 f}{\partial x_i \partial x_j}$ *and* $\frac{\partial^2 f}{\partial x_j \partial x_i}$ *of function f are continuous at* $\mathbf{x}_0$*, then they are equal.*

Thus, Hessian of a sufficiently smooth function is a symmetric matrix.

Theorem 2.3 *All eigenvalues of a symmetric matrix are real.*

Theorem 2.4 *The set of eigenvectors of a symmetric matrix contains an orthonormal basis as a subset.*

Exercise*: Find proofs of these theorems somewhere and try to understand them. (e.g. see M. Kilp. *Algebra I*).

At last, one of the most important results: if a function is twice differentiable, it can be expanded into Taylor's series:

Theorem 2.5
$$f(\mathbf{x}_0 + \mathbf{\Delta x}) = f(\mathbf{x}_0) + \frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{x}}\mathbf{\Delta x} + \frac{1}{2}\mathbf{\Delta x}^T\left(\frac{\partial^2 f(\mathbf{x}_0)}{\partial^2 \mathbf{x}}\right)\mathbf{\Delta x} + o(\|\mathbf{\Delta x}\|^2)$$

Exercise: Prove it.
Hint: Apply the definition of differentiability twice.

The last theorem states, that any sufficiently smooth function, in a sufficiently small neighborhood of $\mathbf{x}$ can be approximated by a second-degree polynomial. It is therefore useful to understand the appearance of a function of the form $f(\mathbf{x}) = c + \mathbf{a}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is symmetric.

Exercise: Consider the function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{H}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{H}$ is symmetric. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be two orthogonal unit eigenvectors of $\mathbf{H}$. Let the corresponding eigenvalues be λ_1 and λ_2. Draw the vectors $\mathbf{v}_1$, $\mathbf{v}_2$ on a plane. Examine how function f behaves on the straight lines defined by these vectors, that is, what are the values of the function at points $t\mathbf{v}_1$, $t\mathbf{v}_2$ ($t \in \mathbb{R}$). Do you see that f is either a convex paraboloid or a surface with a "saddle point". What is the role of the values λ_i?

Exercise: Analyze the appearance of the function $f(\mathbf{x}) = c + \mathbf{a}^T\mathbf{x} + \mathbf{x}^T\mathbf{H}\mathbf{x}$.

3 Notion of a Local Minimum

Let f denote a function $\mathbb{R}^n \to \mathbb{R}$.

Definition Point $\mathbf{x}^*$ is called the *(local) minimum* of f, if there exists a neighborhood $U(\mathbf{x}^*)$ of $\mathbf{x}^*$, such that $f(\mathbf{x}) \geq f(\mathbf{x}^*) \quad \forall \mathbf{x} \in U(\mathbf{x}^*)$.

Theorem 3.1 (Fermat) *Let $\mathbf{x}^*$ be the minimum of f, and let f be differentiable at this point. Then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Exercise: Prove it.

In general the converse does not hold, i.e. gradient being zero does not imply a minimum. It holds, however:

Theorem 3.2 *If f is a convex function, that is*

$$\forall\ \mathbf{x}, \mathbf{y} \quad f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

the condition $\nabla f(\mathbf{x}^*) = \mathbf{0}$ is equivalent to $\mathbf{x}^*$ being the global minimum of f.

Theorem 3.3 *Let $\mathbf{x}^*$ be the minimum of f and let f be twice differentiable at this point. Then $\nabla^2 f(\mathbf{x}^*) \geq 0$ (ie the Hessian of f is positive semidefinite at this point).*

Theorem 3.4 *If f is twice differentiable at point $\mathbf{x}^*$, $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) > 0$, then $\mathbf{x}^*$ is a minimum of f.*

Exercise: Rephrase the last two theorems for the case of $f : \mathbb{R} \to \mathbb{R}$.