1 Continuity

1.1 Continuous Function

A function $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous at $x_0 \in \text{cl} \, D$ if

$$\lim_{x \rightarrow x_0} f(x) = f(x_0),$$

or, equivalently, if for all $\epsilon > 0$ there exists a value $\delta > 0$ such that

$$\|x - x_0\|_2 < \delta \text{ for } x \in D \Rightarrow \|f(x) - f(x_0)\|_2 < \epsilon.$$

A function $f$ is continuous on its domain $D$ if $f$ is continuous for all $x_0 \in \text{cl} \, D$ (note: $D$ will be a closed set for most functions we discuss).

A Special Case ($n = 1$) A function $f : D \subset \mathbb{R} \rightarrow \mathbb{R}$ is continuous at $x_0 \in \text{cl} \, D$ if

$$\lim_{x \uparrow x_0} f(x) = \lim_{x \downarrow x_0} f(x) = f(x_0).$$

For $x_0$ at the boundary of $D$, check either the limit from below or from above.

1.2 Lipschitz Continuity

A function $f : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is Lipschitz continuous on some set $N \subseteq D$ if there exists a constant $L > 0$ such that

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2, \text{ \forall } x, y \in N.$$

If it holds for a neighborhood $N$ of $x \in \text{int} \, D$, $N \subseteq D$, then $f$ is called locally Lipschitz continuous.

Suppose that $f_1$ and $f_2$ are functions from $D \subseteq \mathbb{R}^n$ to $\mathbb{R}^m$, Lipschitz continuous on a set $N \subseteq D$ with constants $L_1 > 0$ and $L_2 > 0$. Then
• $f_1 + f_2$ is Lipschitz continuous on $N$ with a constant $L_1 + L_2$.

• If $m = 1$ and both functions are bounded on $N$ (i.e. $|f_1(x)| \leq M$ and $|f_2(x)| \leq M$ for some $M > 0$), then $f_1 f_2$ is Lipschitz continuous on $N$ with a constant $M(L_1 + L_2)$.

$$|f_1(x)f_2(x) - f_1(y)f_2(y)| \leq |f_1(x)f_2(x) - f_1(y)f_2(x)| + |f_1(y)f_2(x) - f_1(y)f_2(y)|$$
$$= |f_2(x)||f_1(x) - f_1(y)| + |f_1(y)||f_2(x) - f_2(y)|$$
$$\leq M(L_1 + L_2)||x - y||_2.$$

2 Derivatives

2.1 On $\mathbb{R}$

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a real-valued function of a real variable.

\[
\begin{align*}
\{1\text{st derivative}\} & \quad \frac{d\phi}{d\alpha} = \phi'(\alpha) := \lim_{\epsilon \to 0} \frac{\phi(\alpha + \epsilon) - \phi(\alpha)}{\epsilon} \\
\{2\text{nd derivative}\} & \quad \frac{d^2\phi}{d\alpha^2} = \phi''(\alpha) := \lim_{\epsilon \to 0} \frac{\phi'(\alpha + \epsilon) - \phi'(\alpha)}{\epsilon}.
\end{align*}
\]

Chain rule: Suppose that $\alpha$ is a function of $\beta$ (we write $\alpha = \alpha(\beta)$). Then the derivative of $\phi$ w.r.t. $\beta$ is

$$\frac{d\phi(\alpha(\beta))}{d\beta} = \frac{d\phi}{d\alpha} \frac{d\alpha}{d\beta} = \phi'(\alpha)\alpha'(\beta).$$

2.2 On $\mathbb{R}^n$

Let $f : \mathbb{R}^n \to \mathbb{R}$, and $x = (x_1, x_2, \ldots, x_n)^T$ is a column vector.

Frechet differentiability $f$ is differentiable at $x$ if there exists a vector $g \in \mathbb{R}^n$ such that

$$\lim_{y \to 0} \frac{f(x + y) - f(x) - g^Ty}{\|y\|} = 0,$$

where $\| \cdot \|$ is any vector norm of $y$.

Gradient The vector $g$ satisfying Eq. (1) is called the gradient of $f$ at $x$, denoted by $\nabla f(x)$:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$
The partial derivative $\frac{\partial f}{\partial x_i}$ w.r.t. a component $x_i$ is given by setting $y = \epsilon e_i$ for $\epsilon > 0$ and the $i$th unit vector $e_i$ (which has 1 at the $i$th element and 0 at the rest):

$$\frac{\partial f}{\partial x_i} = \lim_{\epsilon \to 0} \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}.$$ 

For a function $f(x, y)$ of two vector variables $x$ and $y$, we use $\nabla_y f(x, y)$ to denote the gradient w.r.t. $y$.

**Jacobian matrix** When $f : \mathbb{R}^n \to \mathbb{R}^m$, then $\nabla f(x)$ is an $n \times m$ matrix: $i$th column is $\nabla f_i(x)$. The Jacobian is $J(x) = (\nabla f(x))^T$ with dimensions $m \times n$.

Ex. $f(x) = Ax$ for $A \in \mathbb{R}^{k \times n}$. What is $\nabla f(x)$? What is $J(x)$?

**Hessian** The matrix of second partial derivatives of $f$ is called the Hessian:

$$\nabla^2 f(x) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}$$

Ex. $f(x_1, x_2) = x_1^2 + x_1x_2$. Then,

$$\nabla f(x_1, x_2) = \begin{bmatrix}
2x_1 + x_2 \\
x_1
\end{bmatrix} \quad \nabla^2 f(x_1, x_2) = \begin{bmatrix}
2 & 1 \\
1 & 0
\end{bmatrix}.$$

Ex. $f(x) = x^T H x$ where $H$ is a symmetric matrix. $\nabla f(x)$ and $\nabla^2 f(x)$? Suppose that $x \in \mathbb{R}^n$ and $H \in \mathbb{R}^{n \times n}$. Note that

$$f(x) = x^T H x = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j H_{ij} = \sum_{i=1}^{n} x_i^2 H_{ii} + \sum_{i \neq j} x_i x_j H_{ij}$$

Therefore,

$$[\nabla f(x)]_k = \frac{\partial f(x)}{\partial x_k} = 2x_k H_{kk} + \sum_{j \neq k} x_j H_{kj} + \sum_{i \neq k} x_i H_{ik}$$

$$= 2x_k H_{kk} + 2 \sum_{j \neq k} x_j H_{kj} \quad (H \text{ is symmetric})$$

$$= 2 \sum_{j=1}^{n} x_j H_{kj} = 2H_{kk} x$$

That is,

$$\nabla f(x) = 2H x$$

3
Check that $\nabla^2 f(x) = 2H$.

$f$ is called differentiable on a domain $D$ if $\nabla f(x)$ exists for all $x \in D$. $f$ is called twice differentiable on a domain $D$ if $\nabla^2 f(x)$ exists for all $x \in D$.

**Classes of Continuously Differentiable Functions** $C^1, C^2, \ldots, C^\infty$ $f$ is $k$-times continuously differentiable if $\nabla^k f(x)$ is a continuous function of $x$. The set of such functions is denoted as $C^k$.

Note that if $f \in C^2$, twice continuously differentiable, then its Hessian is a symmetric matrix, since

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

**Chain Rule** Consider $f : \mathbb{R}^k \to \mathbb{R}$, $x : \mathbb{R}^n \to \mathbb{R}^k$, and a composite function $h(t) = f(x(t))$. The chain rule provides $\nabla h(t)$ as follows,

$$\nabla h(t) = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \nabla x_i(t) = \nabla x(t) \nabla f(x(t)) = J(x(t))^T \nabla f(x(t)).$$

Note that $J(x(t)) \in \mathbb{R}^{k \times n}$ and $\nabla f(x(t)) \in \mathbb{R}^k$, so that $\nabla h(t) \in \mathbb{R}^n$.

Ex. $f(x_1, x_2) = x_1^2 + x_1 x_2$ where $x_1(t_1, t_2) = \sin t_1 + t_2^2$ and $x_2(t_1, t_2) = (t_1 + t_2)^2$. What is $\nabla h(t)$ for $h(t) := f(x(t))$?

$$\nabla h(t) = \sum_{i=1}^2 \frac{\partial f}{\partial x_i} \nabla x_i(t)$$

$$= (2x_1 + x_2) \begin{bmatrix} \cos t_1 \\ 2t_2 \end{bmatrix} + x_1 \begin{bmatrix} 2(t_1 + t_2) \\ 2(t_1 + t_2) \end{bmatrix}$$

$$= (2(\sin t_1 + t_2^2) + (t_1 + t_2)^2) \begin{bmatrix} \cos t_1 \\ 2t_2 \end{bmatrix} + (\sin t_1 + t_2^2) \begin{bmatrix} 2(t_1 + t_2) \\ 2(t_1 + t_2) \end{bmatrix}$$

Ex. $x(t) = At \in \mathbb{R}^k$ for $A \in \mathbb{R}^{k \times n}$ and $t \in \mathbb{R}^n$, $f(x) = \|x\|^2_2 = x^T x$ and $h(t) := f(x(t))$.

(Method 1) $\nabla x(t) = A^T \Rightarrow \nabla h(t) = 2A^T At$

( Method 2) $h(t) = (At)^T (At) = t^T A^T At \Rightarrow \nabla h(t) = 2A^T At$

**References**
