MATH-UA 129: Differentiation

James Pagan

Septemper 2023

Contents

1 Continuity of Sequences and Functions

1.1 Limits of Multivariable Sequences

Limit of a Sequence: Let $\mathbf{x}_r$ for $r \in \mathbb{Z}_{>0}$ be a sequence of vectors in $\mathbb{R}^n$ for $r \in \mathbb{N}$. We say that $\mathbf{x}_r$ *converges* to the vector $\mathbf{L}$ if for all $\epsilon > 0$, there is $N > 0$ such that

$$N < i \implies \|\mathbf{x}_i - \mathbf{L}\| < \epsilon.$$

This is the definition of a limit via metric spaces. Before we can apply the usual limit rules to $\mathbb{R}^n$, we must verify that $\mathbb{R}^n$ is a complete metric space:

Lemma. *$\mathbb{R}^n$ is a complete metric space under the canonical norm.*

Proof. Clearly $\mathbb{R}^n$ is a metric space. Now, let $\mathbf{x}_r = (a_{r1}, \ldots, a_{rn})$ be a Cauchy sequence in $\mathbb{R}^n$; for all $\epsilon > 0$, there is $N > 0$ such that

$$N < i \leq j \implies \|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon.$$

Now, see that for $k \in \{1, \ldots, n\}$,

$$|a_{ik} - a_{jk}| = \sqrt{(a_{ik} - a_{jk})^2} \leq \sqrt{(a_{i1} - a_{j1})^2 + \cdots + (a_{in} - a_{jn})^2} = \|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon,$$

so the sequence a_{rk} is a real Cauchy sequence; as all such sequences converge, we may define $N_1, \ldots, N_n$ and $L_1, \ldots, L_n$ such that

$$\begin{gathered} N_1 < i \implies |a_{i,1} - L_1| < \tfrac{\epsilon}{n}, \\ \vdots \\ N_n < i \implies |a_{i,n} - L_n| < \tfrac{\epsilon}{n}. \end{gathered}$$

Define $\mathbf{L} = (L_1, \ldots, L_n)$. Then for all $\max\{N_1, \ldots, N_n\} < i$, we find that

$$\begin{aligned} \|\mathbf{x}_i - \mathbf{L}\| &= \sqrt{(a_{i,1} - L_1)^2 + \cdots + (a_{i,n} - L_n)^2} \\ &\leq \sqrt{(a_{i,1} - L_1)^2} + \cdots + \sqrt{(a_{i,n} - L_n)^2} \\ &= |a_{i,1} - L_1| + \cdots + |a_{i,n} - L_n| \\ &< \tfrac{\epsilon}{n} + \cdots + \tfrac{\epsilon}{n} \\ &= \epsilon, \end{aligned}$$

so $\mathbf{x}_r$ converges to $\mathbf{L}$, and $\mathbb{R}^n$ is a complete metric space.

1.2 A Brief Topological Excursion: Open Sets

An **open disk** $D_r(\mathbf{x}_0)$ of radius r and center $\mathbf{x_0}$ is the set of all points $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}_0\| < r$. A set $U \subseteq \mathbb{R}^n$ is an **open set** if for all $\mathbf{x}_0 \in U$, there exists $r > 0$ such that $D_r(\mathbf{x}_0) \in U$ — in other words, if it is possible to construct an open disc, no matter how small, around any point in U that lies entirely in U.

Lemma. *Any open disc $D_r(\mathbf{x}_0)$ for $r > 0$ and $\mathbf{x}_0 \in \mathbb{R}^n$ is an open set.*

Proof. Let $\mathbf{y}$ be any point in $D_0(\mathbf{x}_0)$; note that $\mathbf{y}$ satisfies $\|\mathbf{y} - \mathbf{x}_0\| < r$. We wish to construct an open disc centered at $\mathbf{y}$ that lies inside $D_r(\mathbf{x}_0)$.

Let $s = r - \|\mathbf{y} - \mathbf{x}_0\|$. Consider an arbitrary point $\mathbf{z}$ in the the open disc $D_r(\mathbf{y})$ — we have that $\|\mathbf{z} - \mathbf{y}\| < s = r - \|\mathbf{y} - \mathbf{x_0}\|$. Thus, the distance between $\mathbf{z}$ and $\mathbf{x}_0$ satisfies

$$\|\mathbf{z} - \mathbf{x}_0\| \leq \|\mathbf{z} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}_0\| < r - \|\mathbf{y} - \mathbf{x}_0\| + \|\mathbf{y} - \mathbf{x}_0\| = r.$$

Then $\mathbf{z}$ lies in $D_r(\mathbf{x}_0)$; therefore, $D_s \subseteq D_r(x_0)$. Every point $y \in D_r(\mathbf{x}_0)$ is contained within the open disc $D_s(\mathbf{y})$, so $D_r(\mathbf{x}_0)$ is an open set, as desired.

A **neighborhood** of $\mathbf{x} \in \mathbb{R}^n$ is a set that contains an open set that contains $\mathbf{x}$; neighborhoods need not be open sets at all.

A point $x \in \mathbb{R}^n$ is a **boundary point** of $U \subseteq \mathbb{R}^n$ if every neighborhood of x contains a point in U and a point not in U. The set of all boundary points of U is denoted ∂U. We an now define:

1.3 Limits of Multivariable Functions

Let U be an open subset of $\mathbb{R}^n$ and f be a vector-valued function $f : U \to \mathbb{R}^m$, and consider $x \in U$ or $x \in \partial U$. There are two equivalent definitions of limits of multivariable functions taught in this class:

Limit of a Function: We say that $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} = \mathbf{L}$ for $\mathbf{L} \in \mathbb{R}^m$ if for all $\epsilon > 0$, there exists δ such that

$$0 < \|\mathbf{x} - \mathbf{x}_0\| < \delta \implies \|f(\mathbf{x}) - \mathbf{L}\| < \epsilon.$$

Topological Limit of a Function: We say that $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} = \mathbf{L}$ if for all neighborhoods $\mathcal{N}$ of $\mathbf{L}$, there exists a neighborhood U of $\mathbf{x_0}$ such that

$$\mathbf{x} \in U \setminus \mathbf{x}_0 \implies f(\mathbf{x}) \in \mathcal{N}.$$

Continuity: In either definition, we say that f is continuous at $\mathbf{x}_0 \in U$ or $\mathbf{x}_0 \in \partial U$ if $\lim\limits_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$. f *itself* is continuous if f is continuous at all $\mathbf{x}_0 \in U$.

2 Limits on Matricies

2.1 Matrix Norm

Before we can perform calculus on matricies, we must define a metric: the matrix norm. As $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a vector space whose elements are matricies, we can define $\|T\|$ for $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ as follows:

$$\|T\| = \sup\{\|T\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| \leq 1\}.$$

Imposing that $\|\mathbf{x}\| = 1$ yields an identical definition. We have that $T\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \leq \|T\|$ — or more genearlly, that $\|T\mathbf{x}\| \leq \|T\|\|\mathbf{x}\|$. Fascinatingly, the matrix norm satisfies a host of convenient properties:

- **Zero**: If $\|T\| = 0$, then $T\mathbf{x} = 0$ for all $\mathbf{x} \in \mathbb{R}^n$ such $\|\mathbf{x}\| = 1$; we find that for *all* $\mathbf{v} \in \mathbb{R}^n$,

$$T\mathbf{v} = \|\mathbf{v}\|(T\tfrac{\mathbf{v}}{\|\mathbf{v}\|}) = \|\mathbf{v}\|(0) = 0,$$

  so $T = 0$.

- **Positivity**: If $T, S \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that $T \neq S$, then by contraposition, $T - S \neq 0$ implies that $\|T - S\| \neq 0$, so $\|T - S\| > 0$.

- **Additivity**: For matricies $T, S \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, we define define the vector $\mathbf{x} = \sup\{\|(T+S)\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n, \|x\| \leq 1\}$; we have that

$$\|T + S\| = \|(T+S)\mathbf{x}\| \leq \|T\mathbf{x}\| + \|S\mathbf{x}\| \leq \|T\|\|\mathbf{x}\| + \|S\|\|\mathbf{x}\| \leq \|T\| + \|S\|.$$

- **Scalar Multiplication**: For all $\lambda \in \mathbb{R}$, let $\mathbf{x} = \sup\{\|\lambda T\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n, \|x\| \leq 1\}$; then

$$\|\lambda T\| = \|\lambda T\mathbf{x}\| = \lambda\|T\mathbf{x}\| \leq \lambda\|T\|\|\mathbf{x}\| \leq \lambda\|T\|.$$

- **Matrix Multiplication**: If $T \in \mathcal{L}(\mathbb{R}^n), \mathbb{R}^m$ and $S \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^r)$, define the vector $\mathbf{x} = \sup\{\|(TS)\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n, \|x\| \leq 1\}$. Then

$$\|TS\| = \|TS\mathbf{x}\| \leq \|T\|\|S\mathbf{x}\| \leq \|T\|\|S\|\|\mathbf{x}\| \leq \|T\|\|S\|.$$

- **Triangle Inequality**: If $T, S, R \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, then

$$\|T - S\| = \|(T - R) + (R - S)\| \leq \|T - R\| + \|R - S\|.$$

We conclude that the matrix norm is a metric of $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

2.2 Continuity of Linear Maps

Lemma. *If $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, then $\|T\| < \infty$ and T is a uniformly continuous mapping of $\mathbb{R}^n$ into $\mathbb{R}^m$.*

Proof. Define $\mathbf{e}_1, \ldots, \mathbf{e}_n$ as the standard basis of $\mathbb{R}^n$, and let $\mathbf{x} = \sup\{\|T\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n, \|x\| \leq 1\}$ such that $\|T\| = \|T\mathbf{x}\|$ have the form $\mathbf{x} = (x_1, \ldots, x_n)$. Clearly $|x_1|, \ldots, |x_n| \leq 1$, so

$$\begin{aligned}
\|T\| = \|T\mathbf{x}\| &= \|T(x_1\mathbf{e}_1) + \cdots + x_n\mathbf{e}_n\| \\
&\leq |x_1|\|T\mathbf{e}_1\| + \cdots + |x_n|\|T\mathbf{e_n}\| \\
&\leq \|T\mathbf{e}_1\| + \cdots + \|T\mathbf{e}_n\| \\
&< \infty.
\end{aligned}$$

Therefore, for all $\epsilon > 0$, we have that $0 < \|\mathbf{x} - \mathbf{y}\| < \frac{\epsilon}{\|T\|}$ implies that

$$\|T\mathbf{x} - T\mathbf{y}\| = \|T(\mathbf{x} - \mathbf{y})\| \leq \|T\|\|\mathbf{x} - \mathbf{y}\| < \|T\|\frac{\epsilon}{\|T\|} = \epsilon,$$

so T is uniformly continuous.

2.3 A Neat Inequality

Suppose $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{e}_1, \ldots, \mathbf{e}_m$ are the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. Then if $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$,

$$T\mathbf{v} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} v_j \right) \mathbf{e}_i = \left(\sum_{j=1}^{n} a_{1j} v_j, \ldots, \sum_{j=1}^{n} a_{mj} v_j \right).$$

Then via the Cauchy-Schwarz Inequality, we have that for all $\mathbf{v} \in \mathbb{R}^n$,

$$\begin{aligned}
\|T\mathbf{v}\|^2 &= \left(\sum_{j=1}^{n} a_{1j} v_j \right)^2 + \cdots + \left(\sum_{j=1}^{n} a_{mj} v_j \right)^2 \\
&\leq \left(\sum_{j=1}^{n} a_{1j}^2 \right) \left(\sum_{j=1}^{n} v_j^2 \right) + \cdots + \left(\sum_{j=1}^{n} a_{mj}^2 \right) \left(\sum_{j=1}^{n} v_j^2 \right) \\
&= \left(\sum_{j=1}^{n} v_j^2 \right) \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right) = \|\mathbf{v}\|^2 \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right).
\end{aligned}$$

Combining this with the above inequality, we find that if $\mathbf{x} = \sup\{\|T\mathbf{x}\| \mid \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| \leq 1\}$,

$$\|T\| = \|T\mathbf{x}\| \leq \|\mathbf{x}\| \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} \leq \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

2.4 Completeness of Matricies

Lemma. *$\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ equipped with the matrix norm is a complete metric space.*

Proof. Let the sequence $T_1, T_2, \ldots \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ be a Cauchy sequence; declare that for all $\epsilon > 0$, there exists $N > 0$ such that

$$N < i \leq j \implies \|T_i - T_j\| < \epsilon.$$

Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ be the canonical basis of $\mathbb{R}^n$. Then for all $k \in \{1, \ldots, n\}$,

$$N < i \leq j \implies \|T_i \mathbf{e}_k - T_j \mathbf{e}_k\| = \|(T_i - T_j)\mathbf{e}_k\| \leq \|T_i - T_j\| \|\mathbf{e}_k\| = \|T_i - T_j\| < \epsilon,$$

so the sequences of vectors $T_1\mathbf{e}_k, T_2\mathbf{e}_k, \ldots \in \mathbb{R}^m$ for each $k \in \{1, \ldots, n\}$ are Cauchy sequences. Observe that these sequences are the kth column vectors of $T_1, T_2, \ldots \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, and that as $\mathbb{R}^m$ is a complete metric space under the canonical norm, all Cauchy sequences of vectors converge to a unique vector. The coordinates of these vectors — which are the entries of the matricies in the sequence — thus converge.

We now define the limits of the entries of the matricies in our sequence. For notational convenience, define $S = \{(a, b) \in \mathbb{Z}^2 \mid a \in \{1, \ldots, m\}, b \in \{1, \ldots, n\}\}$. For all $(a, b) \in S$, and $i \in \mathbb{Z}_{>0}$, let ${}_it_{ab}$ be the entry in the ath row and bth column of T_i. Then for each $(a, b) \in S$, there exists $L_{ab} \in \mathbb{R}$ such that for all $\epsilon > 0$, there exist nexists a nonnegative integer N_{ab} such that

$$N_{ab} < i \implies |{}_it_{ab} - L_{ab}| < \frac{\epsilon}{\sqrt{mn}}.$$

Define $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ as the matrix with entry L_{ab} in the ath row and bth column for each $(a, b) \in S$. Then $\max\{N_{ab} \mid (a, b) \in S\} < i$ implies that

$$\|T_i - L\| \leq \sqrt{\sum_{a=1}^{m} \sum_{b=1}^{n} ({}_it_{ab} - L_{ab})^2} < \sqrt{\sum_{b=1}^{m} \sum_{a=1}^{n} \left(\frac{\epsilon^2}{mn}\right)} = \sqrt{\epsilon^2} = \epsilon.$$

All Cauchy sequence $T_1, T_2, \ldots \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ thus converges to some matrix L, which implies that $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a complete metric space.

3 Differentiation

3.1 The Jacobian Matrix

Let U be an open subset of $\mathbb{R}^n$, let $f : U :\to \mathbb{R}^m$, and let $\mathbf{x} \in U$. If there exists a linear map $J \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - J\mathbf{h}\|}{\|\mathbf{h}\|} = 0,$$

we say that f is **differentiable** at $\mathbf{x}$ and write $f'(\mathbf{x}) = J$, where J is the **Jacobian matrix** of f at $\mathbf{x}$ — also called the matrix of partial derivatives, the differential, or the total derivative. If f is differentiable at *all* $\mathbf{x} \in U$, we say that f itself is differentiable over U.

Lemma. *The Jacobian matrix is unique.*

Proof. Define f like above. Suppose that for contradiction that there exist two matricies $J \neq K$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - J\mathbf{h}\|}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - K\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

See that $J - K \neq 0$, so $\|J - K\| > 0$. Then there exist d_1 and d_2 such that

$$\begin{aligned}
0 < \|\mathbf{h}\| < \delta_1 &\implies \frac{\|-f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x}) + J\mathbf{h}\|}{\|\mathbf{h}\|} < \frac{\|J-K\|}{2} \\
0 < \|\mathbf{h}\| < \delta_2 &\implies \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - K\mathbf{h}\|}{\|\mathbf{h}\|} < \frac{\|J-K\|}{2}
\end{aligned}$$

For $0 < \|\mathbf{h}\| < \min\{\delta_1, \delta_2\}$, we have that

$$\begin{aligned}
\|J-K\| &= \frac{\|J-K\|}{2} + \frac{\|J-K\|}{2} \\
&> \frac{\|-f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x}) + J\mathbf{h}\|}{\|\mathbf{h}\|} + \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - K\mathbf{h}\|}{\|\mathbf{h}\|} \\
&\geq \frac{\|(-f(\mathbf{x}+\mathbf{h}) + f(\mathbf{x}) + J\mathbf{h}) + (f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - K\mathbf{h})\|}{\mathbf{h}} \\
&= \frac{\|(J-K)\mathbf{h}\|}{\mathbf{h}},
\end{aligned}$$

so $\|J-K\|\|\mathbf{h}\| > \|(J-K)\mathbf{h}\|$, which is our desired contradiction.

As an example, if $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ and $\mathbf{x} \in \mathbb{R}^n$, then the derivative of T at $\mathbf{x}$ is T, as

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|T(\mathbf{x}+\mathbf{h}) - T\mathbf{x} - T\mathbf{h}\|}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{0}\|}{\|\mathbf{h}\|} = 0.$$

It is very intuitive to think of J as an approimation of f at $\mathbf{x}_0$ — namely, that there exists $r(h)$ such that $f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) = J\mathbf{h} - r(\mathbf{h})$ and $\lim\limits_{\mathbf{h}\to\mathbf{0}} \frac{r(\mathbf{h})}{\mathbf{h}} = 0$. This strategy will be exhibited in the following proof:

3.2 Chain Rule

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^k$. If f is differentiable at $\mathbf{x}_0$ and g is differentiable at $f(\mathbf{x}_0)$ — and if $\mathbf{x}_0$ and $f(\mathbf{x}_0)$ are contained within open sets in the domains of f and g respectively — then $g \circ f$ is differentiable at $\mathbf{x}_0$, and*

$$(g \circ f)' = g'(f(\mathbf{x}_0))f'(\mathbf{x}_0).$$

Proof. Let $f'(\mathbf{x}_0) = J$ and $g'(f(\mathbf{x}_0)) = K$. We have that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h}\|}{\|\mathbf{h}\|} = 0 = \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|g(f(\mathbf{x}_0)+\mathbf{h}) - g(f(\mathbf{x}_0)) - K\mathbf{h}\|}{\|h\|}.$$

Define the function $\mathbf{k} = f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0)$; clearly, $\lim\limits_{\mathbf{h}\to\mathbf{0}} \mathbf{k} = 0$. We have that

$$\begin{aligned}
&g(f(\mathbf{x}_0+\mathbf{h})) - g(f(\mathbf{x}_0)) - (KJ)\mathbf{h} \\
=&g(f(\mathbf{x}_0)+\mathbf{k}) - g(f(\mathbf{x}_0)) - K(J\mathbf{h} - f(\mathbf{x_0}+\mathbf{h}) - f(\mathbf{x}_0) + \mathbf{k}) \\
=&g(f(\mathbf{x}_0)+\mathbf{k}) - g(f(\mathbf{x}_0)) - K\mathbf{k} + K(f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h}).
\end{aligned}$$

We now establish bounds for $\|\mathbf{k}\|$:

$$\|\mathbf{k}\| = \|f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h} + J\mathbf{h}\| \le \|\mathbf{h}\| \left(\|J\| + \frac{\|f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h}\|}{\|\mathbf{h}\|} \right)$$

Then using the composition rule for limits,

$$\begin{aligned}
0 &\le \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|g(f(\mathbf{x}_0-\mathbf{h})) - g(f(\mathbf{x}_0)) - (KJ)\mathbf{h}\|}{\|\mathbf{h}\|} \\
&\le \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|g(f(\mathbf{x}_0)+\mathbf{k}) - g(f(\mathbf{x}_0)) - K\mathbf{k}\|}{\|\mathbf{h}\|} + \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|K\|\|f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h}\|}{\|\mathbf{h}\|} \\
&\le \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|g(f(\mathbf{x}_0)+\mathbf{k}) - g(f(\mathbf{x}_0)) - K\mathbf{k}\|}{\|\mathbf{k}\|} \lim_{\mathbf{h}\to\mathbf{0}} \left(\|J\| + \frac{\|f(\mathbf{x}_0+\mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h}\|}{\mathbf{h}} \right) \\
&= (0)(\|J\| + 0) = 0.
\end{aligned}$$

so $g \circ f)'(\mathbf{x}_0) = KJ = g'(f(\mathbf{x}_0))f'(\mathbf{x}_0)$ as required. Yes, that last step of Rudin's is nonrigorous — define it as a piecewise expression equal to 0 whenever $\mathbf{k} = \mathbf{0}$, etc.

3.3 The Partial Derivative

Consider $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$, where U is an open subset of $\mathbb{R}^n$; let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and $\mathbf{e}_1, \ldots, \mathbf{e}_m$ be the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. The **components** of f are the real functions $f_1, \ldots, f_m$ defined by

$$f(\mathbf{x}) = f_1(\mathbf{x})\mathbf{e}_1 + \cdots + f_m(\mathbf{x})\mathbf{e}_m,$$

or equivalently by $f_i(\mathbf{x}) = f(\mathbf{x}) \cdot \mathbf{e}_i$ for each $i \in \{1, \ldots, m\}$. Then for $x \in U$, $i \in \{1, \ldots, m\}$, and $j \in \{1, \ldots, n\}$, we define the **partial derivative** of f_i with respect to x_j as

$$\frac{\partial f_i}{\partial x_j} = \lim_{t \to 0} \frac{f_i(\mathbf{x} + t\mathbf{e}_j) - f_i(\mathbf{x})}{t}.$$

The fact this is a one-variable limit explains why the computation of partial derivatives aligns so closely with differentiation of univarite functions.

Lemma. *The entries of the Jacobian matrix are the partial derivatives: namely, if $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ (where U is open) and f is differentiable at x_0, then the partial derivatives exist and*

$$f'(\mathbf{x}_0)\mathbf{e}_j = \sum_{i=1}^{m} \left(\frac{\partial f_i}{\partial x_j} \right) (\mathbf{x})\mathbf{e}_i$$

Proof. Let j be any integer in the set $\{1, \ldots, n\}$. Since f is differentiable at $\mathbf{x}$,

$$\lim_{t \to 0} \frac{f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0)}{t} - f'(\mathbf{x})\mathbf{e}_j = \lim_{t \to 0} \frac{\|f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0) - f'(\mathbf{x}_0)(t\mathbf{e}_j)\|}{\|t\mathbf{e}_j\|} = 0.$$

Therefore,

$$\begin{aligned}
f'(\mathbf{x})\mathbf{e}_j &= \lim_{t \to 0} \frac{f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0)}{t} \\
&= \lim_{t \to 0} \frac{\sum_{i=1}^{m}(f_1(\mathbf{x}_0 + t\mathbf{e}_j)\mathbf{e}_i) - \sum_{i=1}^{m}(f_i(\mathbf{x}_0)\mathbf{e}_i)}{t} \\
&= \sum_{i=1}^{m} \left(\lim_{t \to 0} \frac{f(\mathbf{x}_0 + t\mathbf{e}_j) - f(\mathbf{x}_0)}{t} \mathbf{e}_i \right) \\
&= \sum_{i=1}^{m} \left(\frac{\partial f_i}{\partial x_j} \right) (\mathbf{x}_0)\mathbf{e}_i,
\end{aligned}$$

as desired.

4 Special Cases

4.1 Real-Valued Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable real-valued function. Then f' is a 1-by-n matrix:

$$f' = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x^n} \end{bmatrix}.$$

The transpose of this matrix is called the **gradient** of f;

$$\nabla f = f'^{\top} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \frac{\partial f}{\partial x_1}\mathbf{e}_1 + \cdots + \frac{\partial f}{\partial x_n}\mathbf{e}_n,$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is the standard basis of $\mathbb{R}^n$. Observe that $f'\mathbf{v} = \nabla f \cdot \mathbf{v}$ for all $\mathbf{v} \in \mathbb{R}^n$ — so the gradient and the derivative of a real-valued function are dual concepts. We can define the derivative of f as a vector ∇f such that

$$\lim_{\mathbf{h}\to 0} \frac{\|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f \cdot \mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

The **directional derivative** of f at $\mathbf{x}$ along a unit vector $\mathbf{v}$ is defined as

$$\nabla_{\mathbf{v}} f = \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t}$$

Observe that $\nabla_{\mathbf{e}_i} f = \frac{\partial f}{\partial x_i} = \nabla f \cdot \mathbf{e}_i$ for all $i \in \{1, \ldots, n\}$. This might lead us to conclude the following lemma:

Lemma. *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$, then $\nabla_{\mathbf{v}} f = \mathbf{v} \cdot (\nabla f)$ for all unit vectors $\mathbf{v}$.*

Proof. Observe that $f'\mathbf{v} = \nabla f \cdot \mathbf{v}$, so we may express the definition of the Jacobian matrix in terms of the gradient of f:

$$\begin{aligned}
\nabla f_{\mathbf{v}} f &= \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} \\
&= \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) - \nabla f(t\mathbf{v})}{t} + \lim_{t\to 0} \frac{\nabla f \cdot (t\mathbf{v})}{t} \\
&= 0 + \lim_{t\to 0} \nabla f \cdot \mathbf{v} \\
&= \nabla f \cdot \mathbf{v},
\end{aligned}$$

as required.

Lemma. *If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$, then the maximum of $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ across all unit vectors $\mathbf{v}$ occurs when $\mathbf{v}$ points in the direction of $\nabla f(\mathbf{x}_0)$.*

Proof. If $\mathbf{v}$ is a unit vector, then the Cauchy-Schwarz Inequality guarantees that

$$\nabla_{\mathbf{v}} f(\mathbf{x}_0) = \mathbf{v} \cdot \nabla f(\mathbf{x}_0) \leq \|\mathbf{v}\| \|\nabla f(\mathbf{x}_0)\| = \|\nabla f(\mathbf{x}_0)\| = \left(\frac{\nabla f(\mathbf{x}_0)}{\|\mathbf{x}_0\|} \right) \cdot \nabla f(\mathbf{x}_0)$$

so the maximum of $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ occurs when $\mathbf{v}$ is the normalization of the gradient vector and points in the direction of $\nabla f(\mathbf{x}_0)$.

More generally, we have that if θ is the angle between the unit vector $\mathbf{v}$ and ∇f, then

$$\nabla_{\mathbf{v}} f = \mathbf{v} \cdot \nabla f = \|\mathbf{v}\| \|\nabla f\| \cos(\theta) = \|\nabla f\| \cos(\theta).$$

The directional derivative oscilates like a sine wave as $\mathbf{v}$ walks around the unit hypersphere.

4.2 Paths and Curves

A continuous function $\mathbf{c} : [a, b] \subseteq \mathbb{R} \to \mathbb{R}^n$ is called a **path** in $\mathbb{R}^n$. The set of all points on the path $\{\mathbf{c}(t) \mid t \in [a, b]\}$ is called a **curve** with endpoints $\mathbf{c}(a)$ and $\mathbf{c}(b)$. The **components** of $\mathbf{c}(t)$ are the real functions $x_1(t), \ldots, x_n(t)$ defined by

$$\mathbf{c}(t) = x_1(t)\mathbf{e}_1 + \cdots + x_n(t)\mathbf{e}_n,$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is the standard basis of $\mathbb{R}^n$. If $\mathbf{c}$ is differentiable at $t \in [a, b]$, observe that the derivative of $\mathbf{c}$ is an n-by-1 matrix called the **velocity vector**: $\mathbf{c}'(t) = \left(\frac{\partial x_1}{\partial t}, \ldots, \frac{\partial x_n}{\partial t} \right)$. Infinitely many paths trace the same curve — some may be differentiable, others not!

Lemma. *Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a C^1 function, and define the level surface S_k as $\{\mathbf{x} \mid f(\mathbf{x}) = k\}$. Then ∇f is normal to the level surface — in particular, if $\mathbf{c}(t)$ is a path in S such that $c(0) = \mathbf{x}$, then $\nabla f(\mathbf{x}) \cdot \mathbf{c}'(t) = 0$.*

Proof. Let $\mathbf{c}$ be a path in S_k such that $\mathbf{c}(0) = \mathbf{x}$. By definition, $f(\mathbf{c}(t)) = k$; the Chain Rule thus yields that

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{c}(0)) = f'(\mathbf{c}(0))\mathbf{c}'(0) = \nabla f(\mathbf{x}) \cdot \mathbf{c}'(0),$$

as desired.

Paths allow us to examine tangent vectors to surfaces via an elegant framework, as exhibited in the above proof.

5 Linear Approximation

5.1 The General Case

The linear approximation of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ at $\mathbf{x}_0 \in \mathbb{R}^n$ with derivative J at $\mathbf{x}_0$ is given by the following equation:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + J(\mathbf{x} - \mathbf{x}_0),$$

where $f'(\mathbf{x}_0)$ is a linear map that sends $\mathbf{x} - \mathbf{x}_0$ from $\mathbb{R}^n$ to $\mathbb{R}^m$. Observe that if we define $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$, then

$$\lim_{\mathbf{x} \to \mathbf{x}_0} \frac{\|f(\mathbf{x}) - f(\mathbf{x}_0) - J(\mathbf{x} - \mathbf{x}_0)\|}{\|\mathbf{x} - \mathbf{x}_0\|} = \lim_{\mathbf{h} \to \mathbf{0}} \frac{\|f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - J\mathbf{h}\|}{\|\mathbf{h}\|} = 0,$$

so the limit of the error of this approximation approaches 0 as $\mathbf{x}$ approaches $\mathbf{x}_0$. In this sense, the derivative J is *precisely* the linear approximation of f at $\mathbf{x}_0$ — isn't it beautiful?

5.2 Special Cases

For real-valued functions: $f : \mathbb{R}^n \to \mathbb{R}$, the linear approximation at $\mathbf{x}_0 \in \mathbb{R}^n$ is called the tangent hyperplane and is typically formulated via the gradient:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0).$$

For implicit surfaces: $f(\mathbf{x}) = k$ for differentiable $f : \mathbb{R}^n \to \mathbb{R}$, the linear approximation is the tangent hyperpla, precisely those vectors tangent to the level set — and thus, normal to the gradient. The equation of the tangent plane at $\mathbf{x}_0 \in \mathbb{R}^n$ is given by

$$\nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = 0.$$

For paths $f : [a, b] \subseteq \mathbb{R} \to \mathbb{R}^n$, the linear approximation at $t_0 \in [a, b]$ is called the tangent line:

$$\mathbf{c}(t) \approx \mathbf{c}(t_0) + \mathbf{c}'(t_0)(t - t_0).$$