# Mathematical Engineering of Deep Learning 

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy
February 28, 2024

## Contents

Preface - DRAFT ..... 3
1 Introduction - DRAFT ..... 1
1.1 The Age of Deep Learning ..... 1
1.2 A Taste of Tasks and Architectures ..... 7
1.3 Key Ingredients of Deep Learning ..... 12
1.4 DATA, Data, data! ..... 17
1.5 Deep Learning as a Mathematical Engineering Discipline ..... 20
1.6 Notation and Mathematical Background ..... 23
Notes and References ..... 25
2 Principles of Machine Learning - DRAFT ..... 27
2.1 Key Activities of Machine Learning ..... 27
2.2 Supervised Learning ..... 32
2.3 Linear Models at Our Core ..... 39
2.4 Iterative Optimization Based Learning ..... 48
2.5 Generalization, Regularization, and Validation ..... 52
2.6 A Taste of Unsupervised Learning ..... 62
Notes and References ..... 72
3 Simple Neural Networks - DRAFT ..... 75
3.1 Logistic Regression in Statistics ..... 75
3.2 Logistic Regression as a Shallow Neural Network ..... 82
3.3 Multi-class Problems with Softmax ..... 86
3.4 Beyond Linear Decision Boundaries ..... 95
3.5 Shallow Autoencoders ..... 99
Notes and References ..... 111
4 Optimization Algorithms - DRAFT ..... 113
4.1 Formulation of Optimization ..... 113
4.2 Optimization in the Context of Deep Learning ..... 120
4.3 Adaptive Optimization with ADAM ..... 128
4.4 Automatic Differentiation ..... 135
4.5 Additional Techniques for First-Order Methods ..... 143
4.6 Concepts of Second-Order Methods ..... 152
Notes and References ..... 164
5 Feedforward Deep Networks - DRAFT ..... 167
5.1 The General Fully Connected Architecture ..... 167
5.2 The Expressive Power of Neural Networks ..... 173
5.3 Activation Function Alternatives ..... 180
5.4 The Backpropagation Algorithm ..... 184
5.5 Weight Initialization ..... 192

## Contents

5.6 Batch Normalization ..... 194
5.7 Mitigating Overfitting with Dropout and Regularization ..... 197
Notes and References ..... 203
6 Convolutional Neural Networks - DRAFT ..... 205
6.1 Overview of Convolutional Neural Networks ..... 205
6.2 The Convolution Operation ..... 209
6.3 Building a Convolutional Layer ..... 216
6.4 Building a Convolutional Neural Network ..... 226
6.5 Inception, ResNets, and Other Landmark Architectures ..... 236
6.6 Beyond Classification ..... 240
Notes and References ..... 247
7 Sequence Models - DRAFT ..... 249
7.1 Overview of Models and Activities for Sequence Data ..... 249
7.2 Basic Recurrent Neural Networks ..... 255
7.3 Generalizations and Modifications to RNNs ..... 265
7.4 Encoders Decoders and the Attention Mechanism ..... 271
7.5 Transformers ..... 279
Notes and References ..... 294
8 Specialized Architectures and Paradigms - DRAFT ..... 297
8.1 Generative Modelling Principles ..... 297
8.2 Diffusion Models ..... 306
8.3 Generative Adversarial Networks ..... 315
8.4 Reinforcement Learning ..... 328
8.5 Graph Neural Networks ..... 338
Notes and References ..... 353
Epilogue - DRAFT ..... 355
A Some Multivariable Calculus - DRAFT ..... 357
A. 1 Vectors and Functions in $\mathbb{R}^{n}$ ..... 357
A. 2 Derivatives ..... 359
A. 3 The Multivariable Chain Rule ..... 362
A. 4 Taylor's Theorem ..... 364
B Cross Entropy and Other Expectations with Logarithms - DRAFT ..... 367
B. 1 Divergences and Entropies ..... 367
B. 2 Computations for Multivariate Normal Distributions ..... 369
Bibliography ..... 399
Index ..... 401

## A Some Multivariable Calculus - DRAFT

This appendix provides key results and notation from multivariable calculus. It is not an exhaustive summary of multi-variable calculus but rather contains the results needed for the contents of the book.

## A. 1 Vectors and Functions in $\mathbb{R}^{n}$

Denote the set of all the real numbers by $\mathbb{R}$ and the real coordinate space of dimension $n$ by $\mathbb{R}^{n}$. Each element of $\mathbb{R}^{n}$ is an $n$ dimensional vector, interpreted as a column of the form

$$
u=\left(u_{1}, \ldots, u_{n}\right)=\left[\begin{array}{lll}
u_{1} & \cdots & u_{n}
\end{array}\right]^{\top}=\left[\begin{array}{c}
u_{1} \\
\vdots \\
u_{n}
\end{array}\right]
$$

The Euclidean norm of $u \in \mathbb{R}^{n}$, measuring the geometric length of $u$ and also known as the $L_{2}$ norm, is

$$
\|u\|_{2}=\sqrt{u^{\top} u}=\left(\sum_{i=1}^{n} u_{i}^{2}\right)^{1 / 2}
$$

Here the scalar $u^{\top} v$ is the inner product between two vectors $u, v \in \mathbb{R}^{n}$. A normalized form of the inner product, called the cosine of the angle between the two vectors, sometimes simply denoted $\cos \theta$, is,

$$
\begin{equation*}
\cos \theta=\frac{u^{\top} v}{\|u\|_{2}\|v\|_{2}} \tag{A.1}
\end{equation*}
$$

The Euclidean norm is a special case of the $L_{p}$ norm which is defined via,

$$
\|u\|_{p}=\left(\sum_{i=1}^{n}\left|u_{i}\right|^{p}\right)^{1 / p}
$$

for $p \geq 1$. When $p$ in $\|\cdot\|_{p}$ is not specified, we interpret $\|\cdot\|$ as the $L_{2}$ norm.
Focusing on the $L_{2}$ norm and the inner product $u^{\top} v$, the Cauchy-Schwartz inequality is,

$$
\begin{equation*}
\left|u^{\top} v\right| \leq\|u\|\|v\| \tag{A.2}
\end{equation*}
$$

where the two sides are equal if and only if $u$ and $v$ are linearly dependent (that is, $u=c v$ for some $c \in \mathbb{R}$ ). Also, the Euclidean distance (or, simply the distance) between $u$ and $v$ is

## A Some Multivariable Calculus - DRAFT

defined as

$$
\|u-v\|=\left(\sum_{i=1}^{n}\left(u_{i}-v_{i}\right)^{2}\right)^{1 / 2}
$$

An important consequence of the Cauchy-Schwartz inequality is that the Euclidean norm satisfies the triangle inequality: For any $u, v \in \mathbb{R}^{n}$,

$$
\begin{equation*}
\|u+v\| \leq\|u\|+\|v\| \tag{A.3}
\end{equation*}
$$

To see this, observe that

$$
\begin{aligned}
\|u+v\|^{2} & =\|u\|^{2}+\|v\|^{2}+2 u^{\top} v \\
& \leq\|u\|^{2}+\|v\|^{2}+2\|u\|\|v\| \\
& =(\|u\|+\|v\|)^{2}
\end{aligned}
$$

Convergence of a sequence of vectors can be defined via scalar converge of the distance. That is, a sequence of vectors $u^{(1)}, u^{(2)}, \ldots$ in $\mathbb{R}^{n}$ is said to converge to a vector $u \in \mathbb{R}^{n}$, denoted via $\lim _{k \rightarrow \infty} u^{(k)}=u$, if

$$
\lim _{k \rightarrow \infty}\left\|u^{(k)}-u\right\|=0
$$

That is, if for every $\varepsilon>0$ there exists an $N_{0}$ such that for all $k \geq N_{0}$,

$$
\left\|u^{(k)}-u\right\|<\varepsilon
$$

Let $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ be an $n$-dimensional multivariate function that maps each vector $u=$ $\left(u_{1}, \ldots, u_{n}\right)^{\top} \in \mathbb{R}^{n}$ to a real number. Then, the function is said to be continuous at $u \in \mathbb{R}^{n}$ if for any sequence $u^{(1)}, u^{(2)}, \ldots$ such that $\lim _{k \rightarrow \infty} u^{(k)}=u$, we have that

$$
\lim _{k \rightarrow \infty} f\left(u^{(k)}\right)=f(u)
$$

Alternatively, $f$ is continuous at $u \in \mathbb{R}^{n}$ if for every $\varepsilon>0$ there exists $\delta>0$ such that

$$
|f(u)-f(v)|<\varepsilon
$$

for every $v \in \mathbb{R}^{n}$ with $\|u-v\|<\delta$. Continuity of $f$ at $u$ implies that the values of $f$ at $u$ and at $v$ can be made arbitrarily close by setting the point $v$ to be arbitrarily close to $u$.

We can extend the above continuity definitions to multivariate vector valued functions of the form $f: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$ that map every $n$ dimensional real-valued vector to an $m$ dimensional real-valued vector. Such functions can be written as

$$
f(u)=\left[\begin{array}{lll}
f_{1}(u) & \cdots & f_{m}(u) \tag{A.4}
\end{array}\right]^{\top}
$$

where $f_{i}: \mathbb{R}^{n} \rightarrow \mathbb{R}$ for each $i=1, \ldots, m$. Then, the function $f$ is called continuous at $u$ if each $f_{i}$ is continuous at $u$. We say that the function $f$ is continuous on a set $\mathcal{U} \subseteq \mathbb{R}^{n}$ if $f$ is continuous at each point in $\mathcal{U}$.

## A. 2 Derivatives

Consider an $n$-dimensional multivariate function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$. The partial derivative $\frac{\partial f(u)}{\partial u_{i}}$ of $f$ with respect $u_{i}$ is the derivative taken with respect to the variable $u_{i}$ while keeping all other variables constant. That is,

$$
\begin{equation*}
\frac{\partial f(u)}{\partial u_{i}}=\lim _{h \rightarrow 0} \frac{f\left(u_{1}, \ldots, u_{i-1}, u_{i}+h, u_{i+1}, \ldots, u_{n}\right)-f(u)}{h} \tag{A.5}
\end{equation*}
$$

Suppose that the partial derivative (A.5) exists for all $i=1, \ldots, n$. Then the gradient of $f$ at $u$, denoted by $\nabla f(u)$ or $\frac{\partial f(u)}{\partial u}$, is a concatenation of the partial derivatives of $f$ with respect to all its variables, and it is expressed as a vector:

$$
\begin{equation*}
\nabla f(u)=\frac{\partial f(u)}{\partial u}=\left[\frac{\partial f(u)}{\partial u_{1}} \cdots \frac{\partial f(u)}{\partial u_{n}}\right]^{\top} \tag{A.6}
\end{equation*}
$$

The gradient $\nabla f(u)$ is a vector capturing the direction of the steepest ascent at $u$. Further, $h\|\nabla f(u)\|$ is the increase in $f$ when moving in that direction for infinitesimal distance $h$.

In some situations, instead of a vector form, variables of the function are represented as a matrix. In that scenario, multivariate functions are of form $f: \mathbb{R}^{n} \times \mathbb{R}^{m} \rightarrow \mathbb{R}$, that is, $f$ maps matrices $U=\left(u_{i, j}\right)$ of dimension $n \times m$ to real values $f(U)$. If the partial derivative $\frac{\partial f(U)}{\partial u_{i}, j}$ exists for all $i=1, \ldots, n$ and $j=1, \ldots, m$, it is convenient to use the notation $\frac{\partial f(U)}{\partial U}$ to denote the collection of the partial derivatives of $f$ with respect to all its variables as a matrix of the same dimension $n \times m$,

$$
\frac{\partial f(U)}{\partial U}=\left[\begin{array}{ccc}
\frac{\partial f(U)}{\partial u_{1,1}} & \cdots & \frac{\partial f(U)}{\partial u_{1, m}}  \tag{A.7}\\
\vdots & \ddots & \vdots \\
\frac{\partial f(U)}{\partial u_{n, 1}} & \cdots & \frac{\partial f(U)}{\partial u_{n, m}}
\end{array}\right] .
$$

## Directional Derivatives

The directional derivative of $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ at $u$ in the direction $v \in \mathbb{R}^{n}$ is the scalar defined by

$$
\nabla_{v} f(u)=\lim _{h \rightarrow 0} \frac{f(u+h v)-f(u)}{h}
$$

The directional derivative generalizes the notion of the partial derivative. In fact, the partial derivative $\frac{\partial f(u)}{\partial u_{i}}$ is the directional derivative at $u$ in the direction of the vector $e_{i}$ which consists of 1 at the $i$-th coordinate and zeros everywhere else. This simply follows from the observation that

$$
\nabla_{e_{i}} f(u)=\lim _{h \rightarrow 0} \frac{f\left(u_{1}, \ldots, u_{i-1}, u_{i}+h, u_{i+1}, \ldots, u_{n}\right)-f(u)}{h}=\frac{\partial f(u)}{\partial u_{i}}
$$

## A Some Multivariable Calculus - DRAFT

As consequence, if the gradient of $f$ exists at $u$, the directional derivative exists in every direction $v$ and we have

$$
\begin{equation*}
\nabla_{v} f(u)=v^{\top} \nabla f(u) \tag{A.8}
\end{equation*}
$$

One way to see (A.8) in the case of continuity of the partial derivatives is via a Taylor's theorem based first-order approximation (see Theorem A.1):

$$
f(u+h v)=f(u)+(h v)^{\top} \nabla f(u)+O\left(h^{2}\right)
$$

where $O\left(h^{k}\right)$ denotes a function such that $O\left(h^{k}\right) / h^{k}$ goes to a constant as $h \rightarrow 0$. Thus,

$$
\frac{f(u+h v)-f(u)}{h}=v^{\top} \nabla f(u)+O(h)
$$

Now take the limit $h \rightarrow 0$ on both the sides to get (A.8).
It is useful to note that the directional derivative $\nabla_{v} f(u)$ is maximum in the direction of the gradient in the sense that for all unit length vectors $v$, the choice $v=\nabla f(u) /\|\nabla f(u)\|$ maximizes $\left\|\nabla_{v} f(u)\right\|$. This is a consequence of the Cauchy-Schwartz inequality (A.2):

$$
\left|\nabla_{v} f(u)\right|=\left|v^{\top} \nabla f(u)\right| \leq\|v\|\|\nabla f(u)\|=\|\nabla f(u)\|
$$

Setting $v=\nabla f(u) /\|\nabla f(u)\|$ achieves the equality.

## Jacobians

The Jacobian is useful for functions of the form $f: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$ as in (A.4) where each $f_{i}$ is a real-valued function of $u$. The Jacobian of $f$ at $u$, denoted by $J_{f}$, is the $m \times n$ matrix defined via

$$
J_{f}(u)=\left[\begin{array}{ccc}
\frac{\partial f_{1}(u)}{\partial u_{1}} & \ldots & \frac{\partial f_{1}(u)}{\partial u_{n}}  \tag{A.9}\\
\vdots & \ddots & \vdots \\
\frac{\partial f_{m}(u)}{\partial u_{1}} & \ldots & \frac{\partial f_{m}(u)}{\partial u_{n}}
\end{array}\right] .
$$

In other words, the $i$-th row of the Jacobian is the gradient $\nabla f_{i}(u)$. In some situations, it is convenient to use the notation $\frac{\partial f(u)}{\partial u}$ to denote the transpose of the Jacobian of $f$ at $u$. That is,

$$
\begin{equation*}
\frac{\partial f(u)}{\partial u}=\left(J_{f}(u)\right)^{\top} \tag{A.10}
\end{equation*}
$$

## Hessians

Returning to functions of the form $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$, to describe the curvature of the function $f$ at a given $u \in \mathbb{R}^{n}$, it is important to consider the second-order partial derivatives at $u$.

These partial derivatives are arranged as an $n \times n$ matrix, called the Hessian and defined by

$$
\nabla^{2} f(u)=\frac{\partial \nabla f(u)}{\partial u}=\left[\begin{array}{cccc}
\frac{\partial^{2} f}{\partial u_{1}^{2}} & \frac{\partial^{2} f}{\partial u_{1} \partial u_{2}} & \cdots & \frac{\partial^{2} f}{\partial u_{1} \partial u_{n}}  \tag{A.11}\\
\frac{\partial^{2} f}{\partial u_{2} \partial u_{1}} & \frac{\partial^{2} f}{\partial u_{2}^{2}} & \cdots & \frac{\partial^{2} f}{\partial u_{2} \partial u_{n}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^{2} f}{\partial u_{n} \partial u_{1}} & \frac{\partial^{2} f}{\partial u_{n} \partial u_{2}} & \cdots & \frac{\partial^{2} f}{\partial u_{n}^{2}}
\end{array}\right]
$$

where $\frac{\partial^{2} f}{\partial u_{i} \partial u_{j}}=\frac{\partial f}{\partial u_{i}}\left(\frac{\partial f}{\partial u_{j}}\right)$. Note that if all the second-order partial derivatives are continuous at $u$, then the Hessian $\nabla^{2} f(u)$ is a symmetric matrix. That is, for all $i, j \in\{1, \ldots, n\}$,

$$
\frac{\partial^{2} f}{\partial u_{i} \partial u_{j}}=\frac{\partial^{2} f}{\partial u_{j} \partial u_{i}}
$$

This result is known as Schwarz's theorem or Clairaut's theorem. Observe that using the Jacobian, we can treat the Hessian as the Jacobian of the gradient vector. That is,

$$
\nabla^{2} f(u)=J_{\nabla f}(u)
$$

Certain attributes of optimization problems are often defined via positive (semi) definiteness of the Hessian $\nabla^{2} f(\theta)$ at $\theta$. In particular, a symmetric matrix $A$ is said to be positive semidefinite if for all $\phi \in \mathbb{R}^{d}$,

$$
\begin{equation*}
\phi^{\top} A \phi \geq 0 . \tag{A.12}
\end{equation*}
$$

Furthermore, $A$ is said to be positive definite if the inequality in (A.12) is strict for all $\phi \in \mathbb{R}^{d} \backslash\{0\}$. Note that the matrix $A$ is called is negative semidefinite (respectively, negative definite) when $-A$ is positive semidefinite (respectively, positive definite).

## Differentiability

A multivariate vector valued function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$ is said to be differentiable at $u \in \mathbb{R}^{n}$ if there is an $m \times n$ dimensional matrix $A$ such that

$$
\lim _{v \rightarrow u}\left(\frac{\|f(u)-f(v)-A(u-v)\|}{\|u-v\|}\right)=0
$$

Here the limit notation $v \rightarrow u$ implies that the limit exists for every sequence $\left\{v^{(k)}: k \geq 1\right\}$ such that $\lim _{k \rightarrow \infty} v^{(k)}=u$. The matrix $A$ is called the derivative. If the function $f$ is differentiable at $u$, then the derivative at $u$ is equal to the Jacobian $J_{f}(u)$. In particular, if $f$ is a real-valued function (that is, $m=1$ ) and differentiable at $u$, then the derivative at $u$ is $\nabla f(u)^{\top}$. If the derivative is continuous on a set $\mathcal{U} \subseteq \mathbb{R}^{n}$, we say that $f$ is continuously differentiable on $\mathcal{U}$, and in that case all the partial derivatives $\frac{\partial f(u)}{\partial u_{i}}$ are continuous on $\mathcal{U}$.

## A. 3 The Multivariable Chain Rule

Consider a multivariate vector valued function $h: \mathbb{R}^{n} \rightarrow \mathbb{R}^{k}$ and a multivariate real-valued function $g: \mathbb{R}^{k} \rightarrow \mathbb{R}$. Suppose that $h$ is differentiable at $u \in \mathbb{R}^{n}$ and $g$ is differentiable at $h(u)=\left[h_{1}(u) \cdots h_{k}(u)\right]^{\top}$. Let $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ be the composition $f=g \circ h$ or $f(u)=g(h(u))$. For each $i=1, \ldots, n$, the multivariate chain rule is,

$$
\frac{\partial f(u)}{\partial u_{i}}=\frac{\partial g(h(u))}{\partial v_{1}} \frac{\partial h_{1}(u)}{\partial u_{i}}+\cdots+\frac{\partial g(h(u))}{\partial v_{k}} \frac{\partial h_{k}(u)}{\partial u_{i}}
$$

where $\frac{\partial g}{\partial v_{i}}$ denotes the partial derivative of $g$ with respect to the $i$-th coordinate. Thus,

$$
\frac{\partial f(u)}{\partial u_{i}}=\left[\begin{array}{lll}
\frac{\partial h_{1}(u)}{\partial u_{i}} & \ldots & \frac{\partial h_{k}(u)}{\partial u_{i}}
\end{array}\right] \nabla g(h(u))
$$

and combining for all $i=1, \ldots, n$,

$$
\nabla f(u)=J_{h}(u)^{\top} \nabla g(h(u))
$$

Now consider the case where $g$ is also a multivariate vector valued function. That is, suppose $h: \mathbb{R}^{n} \rightarrow \mathbb{R}^{k}$ is differentiable at $u$ and $g: \mathbb{R}^{k} \rightarrow \mathbb{R}^{m}$ is differentiable at $h(u)$. Then the composition $f=g \circ h: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$ is a vector valued function with Jacobian,

$$
\begin{equation*}
J_{f}(u)=J_{g}(h(u)) J_{h}(u) \tag{A.13}
\end{equation*}
$$

The expression in (A.13) is called the multivariable chain rule. In terms of the notation (A.10), we may represent the multivariable chain rule as,

$$
\begin{equation*}
\left[\frac{\partial f}{\partial u}\right]^{\top}=\left[\frac{\partial g}{\partial h}\right]^{\top}\left[\frac{\partial h}{\partial u}\right]^{\top}, \quad \text { or } \quad \frac{\partial f}{\partial u}=\frac{\partial h}{\partial u} \frac{\partial g}{\partial h} . \tag{A.14}
\end{equation*}
$$

## The Chain Rule for a Matrix Derivative of an Affine Transformation

Let us focus on the case $y=g(h(u))$ where $h: \mathbb{R}^{n} \rightarrow \mathbb{R}^{k}$ and $g: \mathbb{R}^{k} \rightarrow \mathbb{R}$. Specifically let us assume that $h(\cdot)$ is the affine function $h(u)=W u+b$ where $W \in \mathbb{R}^{k \times n}$ and $b \in \mathbb{R}^{k}$. That is,

$$
y=g(z), \quad \text { with } \quad z=W u+b
$$

We are often interested in the derivative of the scalar output $y$ with respect to the matrix $W=\left[w_{i, j}\right]$. This is denoted via $\frac{\partial y}{\partial W}$ as in (A.7).

It turns out that we can represent this matrix derivative as the outer product,

$$
\begin{equation*}
\frac{\partial y}{\partial W}=\frac{\partial y}{\partial z} u^{\top} \tag{A.15}
\end{equation*}
$$

where $\frac{\partial y}{\partial z}$ is the gradient of $g(\cdot)$ evaluated at $z$.
To see (A.15) denote the columns of $W$ via $w_{(1)}, \ldots, w_{(n)}$, each an element of $\mathbb{R}^{k}$, and observe that,

$$
z=b+\sum_{i=1}^{n} u_{i} w_{(i)}
$$

We may now observe that Jacobian transposed $\partial z / \partial w_{(i)}$ is $u_{i} I$, where $I$ is the $k \times k$ identity matrix. Hence now, using (A.13), we have,

$$
\frac{\partial y}{\partial w_{(i)}}=\frac{\partial z}{\partial w_{(i)}} \frac{\partial y}{\partial z}=u_{i} \frac{\partial y}{\partial z}
$$

Now we can construct $\frac{\partial y}{\partial W}$ column by column,

$$
\frac{\partial y}{\partial W}=\left[\frac{\partial y}{\partial w_{(1)}} \cdots \frac{\partial y}{\partial w_{(n)}}\right]=\left[u_{1} \frac{\partial y}{\partial z} \cdots u_{n} \frac{\partial y}{\partial z}\right]=\frac{\partial y}{\partial z} u^{\top}
$$

## Jacobian Vector Products and Vector Jacobian Products

Let $f=\left(f_{1}, \ldots, f_{m}\right)=h_{L} \circ h_{L-1} \circ \cdots \circ h_{1}$ be a composition of $L$ differentiable functions $h_{1}, h_{2}, \ldots, h_{L}$ such that $h_{\ell}: \mathbb{R}^{m_{\ell-1}} \rightarrow \mathbb{R}^{m_{\ell}}$ where $m_{0}, m_{1}, \ldots, m_{L}$ are positive integers with $m_{0}=n$ and $m_{L}=m$.

Further, to simplify the notation, for each $\ell=1, \ldots, L$, let

$$
g_{\ell}(u)=h_{\ell}\left(h_{\ell-1}\left(\cdots\left(h_{1}(u)\right) \cdots\right)\right)
$$

Then, $g_{L}(u)=f(u)$ and by recursive application of (A.13), we obtain

$$
\begin{equation*}
J_{f}(u)=J_{h_{L}}\left(g_{L-1}(u)\right) J_{h_{L-1}}\left(g_{L-2}(u)\right) \cdots J_{h_{1}}(u) \tag{A.16}
\end{equation*}
$$

Note that from the definition of the Jacobian, the $j$-th column of $J_{f}(u)$ is the $m$ dimensional vector

$$
\frac{\partial f(u)}{\partial u_{j}}=\left(\frac{\partial f_{1}(u)}{\partial u_{j}}, \ldots, \frac{\partial f_{m}(u)}{\partial u_{j}}\right)=J_{f}(u) e_{j}
$$

where $e_{j}$ is the $j$-th unit vector of appropriate dimension. Therefore, using (A.16), for each $j=1, \ldots, n$,

$$
\begin{equation*}
\frac{\partial f(u)}{\partial u_{j}}=J_{h_{L}}\left(g_{L-1}(u)\right)\left[J_{h_{L-1}}\left(g_{L-2}(u)\right)\left[\cdots\left[J_{h_{1}}(u) e_{j}\right] \cdots\right]\right] \tag{A.17}
\end{equation*}
$$

That is, for each $j=1, \ldots, n, \frac{\partial f(u)}{\partial u_{j}}$ can be obtained by recursively computing the Jacobian vector product given by

$$
v_{\ell}:=J_{h_{\ell}}\left(g_{\ell-1}(u)\right) v_{\ell-1},
$$

for $\ell=1, \ldots, L$, starting with $v_{0}=e_{j}$ and $g_{0}(u)=u$.
On the other hand, since the $i$-th row of $J_{f}(u)$ is the gradient $\nabla f_{i}(u)$, we have

$$
\begin{align*}
\nabla f_{i}(u) & =e_{i}^{\top} J_{f}(u) \\
& =\left[\cdots\left[\left[e_{i}^{\top} J_{h_{L}}\left(g_{L-1}(u)\right)\right] J_{h_{L-1}}\left(g_{L-2}(u)\right)\right] \cdots\right] J_{h_{1}}(u) \tag{A.18}
\end{align*}
$$

## A Some Multivariable Calculus - DRAFT

That is, for each $i=1, \ldots, m, \nabla f_{i}(u)$ can be obtained by recursively computing the vector Jacobian product given by

$$
v_{\ell}^{\top}:=v_{\ell-1}^{\top} J_{h_{L-\ell+1}}\left(g_{L-\ell}(u)\right),
$$

for $\ell=1, \ldots, L$, starting with $v_{0}=e_{i}$ and $g_{0}(u)=u$.

## A. 4 Taylor's Theorem

Once again consider a multivariate real-valued function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$. If all the $k$-order derivatives of $f$ are continuous at a point $u \in \mathbb{R}^{n}$, then Taylor's theorem offers an approximation for $f$ within a neighborhood of $u$ in terms of these derivatives. We are particularly interested in cases where $k=1$ and $k=2$ as they are crucial in implementation of, respectively, the first-order and the second-order optimization methods. It is easy to understand the theorem when the function $f$ is univariate. Hence we start with the univariate case and then move to the general multivariate case. We omit the proof of Taylor's theorem as it is a well known result that can be found in any standard multivariate calculus textbook.

## Univariate Case

Suppose that $n=1$, that is, $f$ is a univariate real-valued function. We say that $f$ is $k$-times continuously differentiable on an open interval $\mathcal{U} \subseteq \mathbb{R}$ if $f$ is $k$ times differentiable at every point on $\mathcal{U}$ (i.e., the $k$-th order derivative $\frac{\mathrm{d}^{k} f(u)}{\mathrm{d} u^{k}}$ exists for all $u \in \mathcal{U}$ ) and $\frac{\mathrm{d}^{k} f(u)}{\mathrm{d} u^{k}}$ is continuous on $\mathcal{U}$. If $k=0$, we interpret $\frac{\mathrm{d}^{k} f(u)}{\mathrm{d} u^{k}}$ simply as $f(u)$.

Theorem A. 1 (Taylor's Theorem in $\mathbb{R}$ ). Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be $k$-times continuously differentiable on an open interval $\mathcal{U} \subseteq \mathbb{R}$. Then, for any $u, v \in \mathcal{U}$,

$$
\begin{equation*}
f(u)=\sum_{i=0}^{k} \frac{(u-v)^{i}}{i!} \frac{\mathrm{d}^{i} f(v)}{\mathrm{d} u^{i}}+O\left(|u-v|^{k+1}\right) \tag{A.19}
\end{equation*}
$$

The polynomial,

$$
P_{k}(u)=\sum_{i=0}^{k} \frac{(u-v)^{i}}{i!} \frac{\mathrm{d}^{i} f(v)}{\mathrm{d} u^{i}}
$$

appeared in (A.19) is called $k$-th order Taylor polynomial. Since the remainder

$$
R_{k}(u)=f(u)-P_{k}(u) \longrightarrow 0, \quad \text { as } x \rightarrow a
$$

$f(u)$ is approximately equal to $P_{k}(u)$ for $u$ within a small neighborhood of $a$. Particularly, for a point $u$ near $v, P_{1}(u)$ is linear approximation of $f(u)$ and $P_{2}(u)$ is quadratic approximation of $f(u)$.

## Multivariate Case

Now consider the multivariate case, that is, $f$ is a multivariate real-valued function. In order to state Taylor's theorem for this case, we need some new notion that is relevant only here.

An $n$-tuple $\alpha=\left(\alpha_{1}, \ldots, \alpha_{n}\right)$ is called multi-index if each $\alpha_{i}$ is a non-negative integer. For a multi-index $\alpha$, let

$$
|\alpha|=\sum_{i=1}^{n} \alpha_{i}, \quad \alpha!=\alpha_{1}!\cdots \alpha_{n}!, \quad \text { and } \quad u^{\alpha}=u_{1}^{\alpha_{1}} \cdots u_{n}^{\alpha_{n}}
$$

for any $u \in \mathbb{R}^{n}$. Then, the higher order partial derivatives are expressed as

$$
D^{\alpha} f(u)=\frac{\partial^{|\alpha|} f(u)}{\partial u_{1}^{\alpha_{1}} \cdots \partial u_{n}^{\alpha_{n}}}
$$

We say that $f$ is $k$-times continuously differentiable on an open set $\mathcal{U} \subseteq \mathbb{R}^{n}$ if all the higher order partial derivatives $D^{\alpha} f(u)$ exists and are continuous on $\mathcal{U}$ for all multi-index $\alpha$ such that $|\alpha| \leq k$.

Theorem A. 2 (Taylor's Theorem in $\mathbb{R}^{n}$ ). Let $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ be a $k$-times continuously differentiable on an open set $\mathcal{U} \subseteq \mathbb{R}^{n}$. Then, for any $u, v \in \mathcal{U}$,

$$
\begin{equation*}
f(u)=\sum_{\alpha:|\alpha| \leq k} D^{\alpha} f(v) \frac{(u-v)^{\alpha}}{\alpha!}+O\left(\|u-v\|^{k+1}\right) \tag{A.20}
\end{equation*}
$$

The polynomial,

$$
P_{k}(u)=\sum_{\alpha:|\alpha| \leq k} D^{\alpha} f(v) \frac{(u-v)^{\alpha}}{\alpha!}
$$

is called $k$-th order Taylor's polynomial. In particular,

$$
\begin{equation*}
P_{1}(u)=\sum_{\alpha:|\alpha| \leq 1} D^{\alpha} f(v) \frac{(u-v)^{\alpha}}{\alpha!}=f(v)+(u-v)^{\top} \nabla f(a) \tag{A.21}
\end{equation*}
$$

for $u$ near $v$, provides linear approximation, also called first-order Taylor's approximation, to $f(u)$, while

$$
\begin{align*}
P_{2}(u) & =\sum_{\alpha:|\alpha| \leq 2} D^{\alpha} f(v) \frac{(u-v)^{\alpha}}{\alpha!} \\
& =f(v)+(u-v)^{\top} \nabla f(v)+\frac{1}{2}(u-v)^{\top} \nabla^{2} f(v)(u-v) \tag{A.22}
\end{align*}
$$

provides quadratic approximation, also called second-order Taylor's approximation, to $f(u)$.

## Linear Approximation with Jacobians and Hessians

Consider a differentiable function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$ with the $m \times n$ Jacobian $J_{f}(\cdot)$. Then with Theorem (A.2) we may construct a first order linear approximation to $f(\cdot)$ around any $u_{0} \in \mathbb{R}^{n}$,

$$
\begin{equation*}
\tilde{f}(u)=f\left(u_{0}\right)+J_{f}\left(u_{0}\right)\left(u-u_{0}\right) \tag{A.23}
\end{equation*}
$$

where $\tilde{f}(u) \approx f(u)$.

## A Some Multivariable Calculus - DRAFT

Now consider a twice differentiable $g: \mathbb{R}^{n} \rightarrow \mathbb{R}$ with gradient $\nabla g(\cdot)$ and Hessian matrix $\nabla^{2} g(\cdot)$. We can set $f(u)=\nabla g(u)$ with $f: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$. Since the Hessian of $g(\cdot)$ is the Jacobian of $f(\cdot)$, from (A.23) we obtain a first order linear approximation for the gradient around $u_{0} \in \mathbb{R}^{n}$,

$$
\begin{equation*}
\widetilde{\nabla} g(u)=\nabla g\left(u_{0}\right)+\nabla^{2} g\left(u_{0}\right)\left(u-u_{0}\right) \tag{A.24}
\end{equation*}
$$

where $\widetilde{\nabla} g(u) \approx \nabla g(u)$.

