Analysis in R^n

- Convergence rate

Estimating the order r:

$$r \approx \frac{\log \dfrac{x_{k+1} - x_k}{x_k - x_{k-1}}}{\log \dfrac{x_k - x_{k-1}}{x_{k-1} - x_{k-2}}}.$$

Assume $\mathbf{x}_k \to \mathbf{x}^*$. We say that the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^*$ R-linearly if

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le e_k$$

R-linear

and $\{e_k\}$ converges to 0 Q-linearly.

Remedies the issue when $\lim_{k\to\infty} \dfrac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|^r}$ does not exist.

Example: $x_k = \begin{cases} 1 + 2^{-k}, & k \text{ even}, \\ 1, & k \text{ odd}. \end{cases}$

Analysis in R^n

- Continuity

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *continuous* at $\mathbf{x} \in \operatorname{dom} f$ if for all $\varepsilon > 0$ there exists a δ such that

$$\mathbf{y} \in \operatorname{dom} f,\ \|\mathbf{y} - \mathbf{x}\|_2 \leq \delta \Rightarrow \|f(\mathbf{y}) - f(\mathbf{x})\|_2 \leq \varepsilon.$$

Continuity can be described in terms of limits: whenever the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ in $\operatorname{dom} f$ converges to a point $\mathbf{x} \in \operatorname{dom} f$, the sequence $f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots$ converges to $f(\mathbf{x})$, *i.e.*,

$$\lim_{i \to \infty} f(\mathbf{x}_i) = f(\lim_{i \to \infty} \mathbf{x}_i).$$

A function f is continuous if it is continuous at every point in its domain.

Analysis in R^n

- Minimum and minimal

A point $\mathbf{x}^*$ is called a *minimum point* of a function $f(\mathbf{x})$ if

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in \operatorname{dom} f.$$

Accordingly, $f(\mathbf{x}^*)$ is called the *minimum value* of f.

$\mathbf{x}^*$ is called a *minimal point* of f if for sufficiently small $\varepsilon > 0$

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^*) \cap \operatorname{dom} f.$$

Accordingly, $f(\mathbf{x}^*)$ is called the *minimal value* of f.

Analysis in R^n

- Closedness

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *closed* if, for each $\alpha \in \mathbb{R}$, the sublevel set

$$\{\mathbf{x} \in \operatorname{dom} f | f(\mathbf{x}) \leq \alpha\}$$

is closed. This is equivalent to the condition that the epigraph of f,

$$\operatorname{epi} f = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} | \mathbf{x} \in \operatorname{dom} f, f(\mathbf{x}) \leq t\},$$

is closed.

If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, and $\operatorname{dom} f$ is closed, then f is closed. If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, with $\operatorname{dom} f$ open, then f is closed iff f converges to ∞ along every sequence converging to a boundary point of $\operatorname{dom} f$. In other words, if $\lim_{i\to\infty} \mathbf{x}_i = \mathbf{x} \in \partial(\operatorname{dom} f)$, with $\mathbf{x}_i \in \operatorname{dom} f$, we have $\lim_{i\to\infty} f(\mathbf{x}_i) = \infty$.

Examples: $f(x) = x \log x$ with $\operatorname{dom} f = \mathbb{R}_{++}$; $f(x) = -\log x$ with $\operatorname{dom} f = \mathbb{R}_{++}$; $f(x) = \begin{cases} x \log x, & x > 0 \\ 0, & x = 0, \end{cases}$ with $\operatorname{dom} f = \mathbb{R}_+$

Analysis in R^n

- Derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$, and $\mathbf{x} \in (\operatorname{dom} f)^\circ$. If there exists a matrix $\mathbf{J}$ such that

$$\lim_{\mathbf{z}\in\operatorname{dom} f, \mathbf{z}\neq\mathbf{x}, \mathbf{z}\to\mathbf{x}} \frac{\|f(\mathbf{z}) - f(\mathbf{x}) - \mathbf{J}(\mathbf{z}-\mathbf{x})\|_2}{\|\mathbf{z}-\mathbf{x}\|_2} = 0,$$

for all choice of sequence $\{\mathbf{z}\} \subset \operatorname{dom} f$, then f is said to be differentiable at $\mathbf{x}$ and denote $Df(\mathbf{x}) = \mathbf{J}$. Let $\mathbf{z} = \mathbf{x} + t\mathbf{e}_i$ and let $t \to 0$. Then

$$\begin{aligned}
\lim_{\mathbf{z}\in\operatorname{dom} f, \mathbf{z}\neq\mathbf{x}, \mathbf{z}\to\mathbf{x}} \frac{\|f(\mathbf{z}) - f(\mathbf{x}) - \mathbf{J}(\mathbf{z}-\mathbf{x})\|_2}{\|\mathbf{z}-\mathbf{x}\|_2} &= \lim_{t\to 0} \frac{\|f(\mathbf{x}+t\mathbf{e}_i) - f(\mathbf{x}) - t\mathbf{J}\mathbf{e}_i\|_2}{|t|} \\
&= \lim_{t\to 0} \left\| \frac{f(\mathbf{x}+t\mathbf{e}_i) - f(\mathbf{x})}{t} - \mathbf{J}\mathbf{e}_i \right\|_2 \\
&= \left\| \frac{\partial f(\mathbf{x})}{\partial x_i} - \mathbf{J}\mathbf{e}_i \right\|_2 .
\end{aligned}$$

Therefore, the i-th column of $\mathbf{J}$ is $\frac{\partial f(\mathbf{x})}{\partial x_i}$. Thus $\mathbf{J} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}^T} = \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j} \right)$. Jacobian

Analysis in R^n

- Gradient

When f is real-valued (*i.e.*, $f : \mathbb{R}^n \to \mathbb{R}$) the derivative $Df(\mathbf{x})$ is a $1 \times n$ *row* vector. Its transpose is called the *gradient* of the function:

$$\nabla f(\mathbf{x}) = Df(\mathbf{x})^T.$$

Its components are the partial derivatives of f:

$$\nabla f(\mathbf{x})_i = \frac{\partial f(\mathbf{x})}{\partial x_i}, \quad i = 1, ..., n.$$

The first-order approximation of f at a point $\mathbf{x} \in (\operatorname{dom} f)^\circ$ can be expressed as

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{z} - \mathbf{x}).$$

Examples: $f(\mathbf{x}) = (1/2)\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^\top\mathbf{x} + r$, $f(\mathbf{X}) = \log\det\mathbf{X}$, with $\operatorname{dom} f = \mathbf{S}^n_{++}$

What if $\mathbf{P}$ is not symmetric?

Analysis in R^n

- Gradient of $f(\mathbf{X}) = \log\det\mathbf{X}$, with $\operatorname{dom} f = \mathbb{S}^n_{++}$

Let $\mathbf{Z} \in \mathbb{S}^n_{++}$ be close to $\mathbf{X}$, $\Delta\mathbf{X} = \mathbf{Z} - \mathbf{X}$, and λ_i be the ith eigenvalue of $\mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2}$. We have

$$
\begin{aligned}
\log\det\mathbf{Z} &= \log\det(\mathbf{X}+\Delta\mathbf{X}) \\
&= \log\det\left(\mathbf{X}^{1/2}(\mathbf{I}+\mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}\right) \\
&= \log\det\mathbf{X} + \log\det(\mathbf{I}+\mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2}) \\
&= \log\det\mathbf{X} + \sum_{i=1}^{n}\log(1+\lambda_i) \approx \log\det\mathbf{X} + \sum_{i=1}^{n}\lambda_i \\
&= \log\det\mathbf{X} + \operatorname{tr}(\mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2}) \\
&= \log\det\mathbf{X} + \operatorname{tr}(\mathbf{X}^{-1}\Delta\mathbf{X}) \\
&= \log\det\mathbf{X} + \operatorname{tr}(\mathbf{X}^{-1}(\mathbf{Z}-\mathbf{X})).
\end{aligned}
$$

Hence $\nabla f(\mathbf{X}) = \mathbf{X}^{-1}$.

Analysis in R^n

- Chain rule

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in \operatorname{dom} f$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(\mathbf{x}) \in (\operatorname{dom} g)^\circ$. Define the composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(\mathbf{z}) = g(f(\mathbf{z}))$. Then h is differentiable at $\mathbf{x}$, with derivative

$$Dh(\mathbf{x}) = Dg(f(\mathbf{x}))Df(\mathbf{x}).$$

You may check the dimensions of the matrices in order to determine the order of multiplication.

Examples: suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(\mathbf{x}) = g(f(\mathbf{x}))$, then

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

Composition with affine function: Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable, $\mathbf{A} \in \mathbb{R}^{n\times p}$, and $\mathbf{b} \in \mathbb{R}^n$. Define $g : \mathbb{R}^p \to \mathbb{R}^m$ as $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$, with $\operatorname{dom} g = \{\mathbf{x} | \mathbf{A}\mathbf{x} + \mathbf{b} \in \operatorname{dom} f\}$. The derivative of g is $Dg(\mathbf{x}) = Df(\mathbf{A}\mathbf{x} + \mathbf{b})\mathbf{A}$.

When f is real-valued (*i.e.*, $m = 1$),

$$\nabla g(\mathbf{x}) = \mathbf{A}^\top \nabla f(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

Analysis in R^n

- Gradient of $f(\mathbf{x}) = \log \sum_{i=1}^{m} \exp\left(\mathbf{a}_i^\top \mathbf{x} + b_i\right)$

It is the composition of the affine function $\mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rows $\mathbf{a}_1^T, ..., \mathbf{a}_m^T$, and the function $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(\mathbf{y}) = \log\left(\sum_{i=1}^m \exp y_i\right)$. Simple differentiation shows that

$$\nabla g(\mathbf{y}) = \frac{1}{\sum_{i=1}^m \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix},$$

so by the composition formula we have

$$\nabla f(\mathbf{x}) = \frac{1}{\mathbf{1}^T \mathbf{z}} \mathbf{A}^T \mathbf{z}$$

where $z_i = \exp\left(\mathbf{a}_i^T \mathbf{x} + b_i\right), i = 1, \ldots, m.$

Analysis in R^n

- Second derivative

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in (\operatorname{dom} f)^\circ$. Then $Df(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$. Suppose $Df(\mathbf{x})$ is also differentiable, then $D^2 f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ and

$$D^2 f(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \frac{\partial (Df(\mathbf{x}))^T}{\partial \mathbf{x}^T} = \left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right) = \mathbf{H}.$$

Hessian

The second-order approximation of f at a point $\mathbf{x} \in (\operatorname{dom} f)^\circ$ can be expressed as

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{z} - \mathbf{x}) + \frac{1}{2} (\mathbf{z} - \mathbf{x})^T \nabla^2 f(\mathbf{x}) (\mathbf{z} - \mathbf{x}).$$

Examples: $f(\mathbf{x}) = (1/2)\mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^\top \mathbf{x} + r$, $f(\mathbf{X}) = \log \det \mathbf{X}$, with $\operatorname{dom} f = \mathbf{S}^n_{++}$

Analysis in R^n

- Second derivative of $f(\mathbf{X}) = \log\det\mathbf{X}$, with $\operatorname{dom} f = \mathbb{S}^n_{++}$

For $\mathbf{Z} \in \mathbb{S}^n_{++}$ near $\mathbf{X} \in \mathbb{S}^n_{++}$, and $\Delta\mathbf{X} = \mathbf{Z} - \mathbf{X}$, we have

$$\begin{aligned}
\mathbf{Z}^{-1} = (\mathbf{X} + \Delta\mathbf{X})^{-1} &= \left(\mathbf{X}^{1/2}(\mathbf{I} + \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}\right)^{-1} \\
&= \mathbf{X}^{-1/2}(\mathbf{I} + \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})^{-1}\mathbf{X}^{-1/2} \\
&\approx \mathbf{X}^{-1/2}(\mathbf{I} - \mathbf{X}^{-1/2}\Delta\mathbf{X}\mathbf{X}^{-1/2})\mathbf{X}^{-1/2} \\
&= \mathbf{X}^{-1} - \mathbf{X}^{-1}\Delta\mathbf{X}\mathbf{X}^{-1}.
\end{aligned}$$

The Hessian of f at $\mathbf{X}$ is a fourth-order tensor $\mathcal{T}$. The above shows that $\mathcal{T}(\Delta\mathbf{X}) = -\mathbf{X}^{-1}\Delta\mathbf{X}\mathbf{X}^{-1}$. The (i,j,k,l)-th entry of $\mathcal{T}$ is

$$\begin{aligned}
\mathcal{T}_{ijkl} = \mathbf{e}_i^T[-\mathbf{X}^{-1}(\mathbf{e}_k\mathbf{e}_l^T)\mathbf{X}^{-1}]\mathbf{e}_j &= -(\mathbf{e}_i^T\mathbf{X}^{-1}\mathbf{e}_k)(\mathbf{e}_l^T\mathbf{X}^{-1}\mathbf{e}_j) \\
&= -(\mathbf{X}^{-1})_{ik}(\mathbf{X}^{-1})_{lj}.
\end{aligned}$$

The second-order approximation of f near $\mathbf{X}$ is:

$$f(\mathbf{Z}) \approx f(\mathbf{X}) + \operatorname{tr}(\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})) - (1/2)\operatorname{tr}(\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})\mathbf{X}^{-1}(\mathbf{Z} - \mathbf{X})).$$

Analysis in R^n

- Chain rules for second derivative

Composition with scalar function. Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h(\mathbf{x}) = g(f(\mathbf{x}))$. Simply working out the partial derivatives yields

$$\nabla^2 h(\mathbf{x}) = g'(f(\mathbf{x}))\nabla^2 f(\mathbf{x}) + g''(f(\mathbf{x}))\nabla f(\mathbf{x})\nabla f(\mathbf{x})^T.$$

Composition with affine function. Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times m}$, and $\mathbf{b} \in \mathbb{R}^n$. Define $g : \mathbb{R}^m \to \mathbb{R}$ by $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then we have

$$\nabla^2 g(\mathbf{x}) = \mathbf{A}^\top \nabla^2 f(\mathbf{A}\mathbf{x} + \mathbf{b})\mathbf{A}.$$

Analysis in R^n

- Second derivative of $f(\mathbf{x}) = \log \sum_{i=1}^{m} \exp\left(\mathbf{a}_i^\top \mathbf{x} + b_i\right)$

$$\nabla g(\mathbf{y}) = \frac{1}{\sum_{i=1}^{m} \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix},$$

$$\nabla^2 g(\mathbf{y}) = \operatorname{diag}\left(\nabla g(\mathbf{y})\right) - \nabla g(\mathbf{y}) \nabla g(\mathbf{y})^T.$$

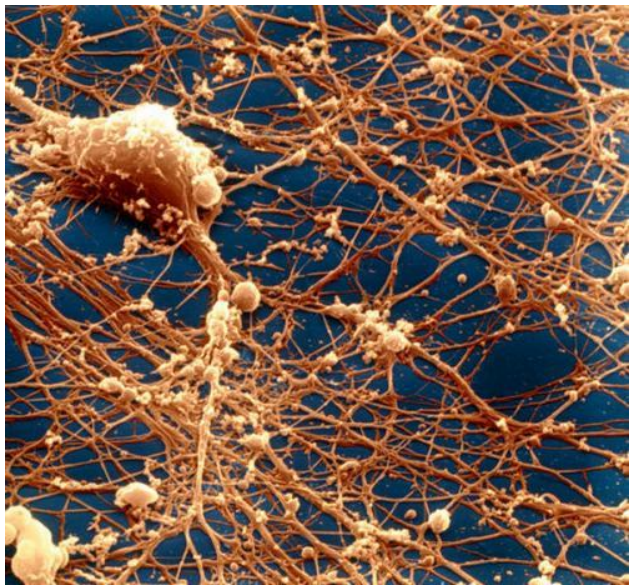
By the composition formula we have

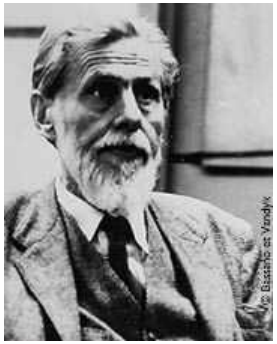
$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^\top \left(\frac{1}{\mathbf{1}^T \mathbf{z}} \operatorname{diag}(\mathbf{z}) - \frac{1}{(\mathbf{1}^T \mathbf{z})^2} \mathbf{z}\mathbf{z}^T \right) \mathbf{A},$$

where $z_i = \exp(\mathbf{a}_i^\top \mathbf{x} + b_i), i = 1, ..., m.$

Analysis in R^n

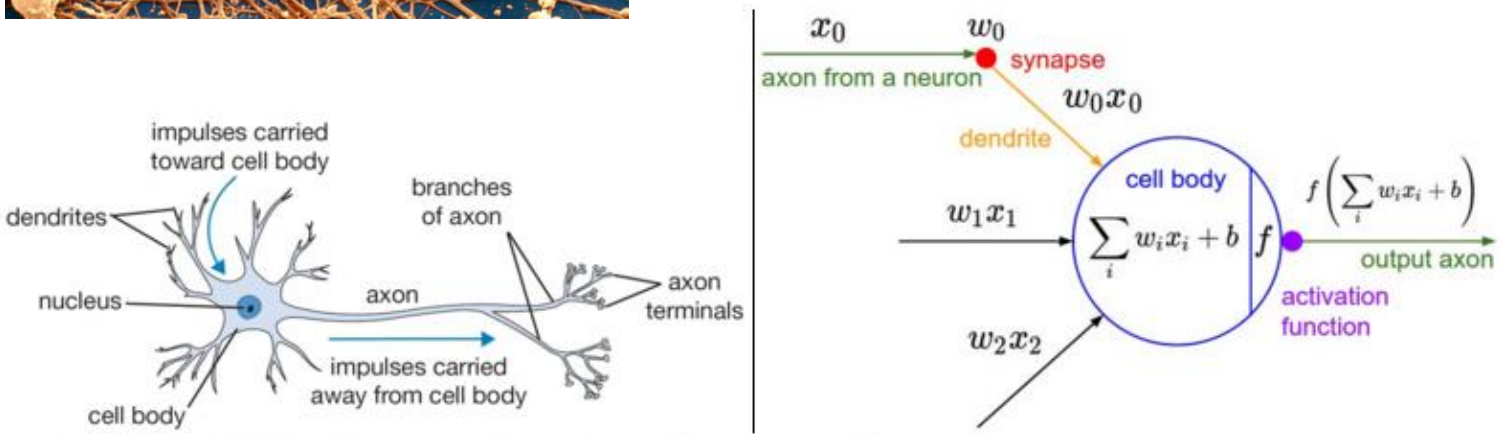
- Deep neural networks





Warren McCulloch

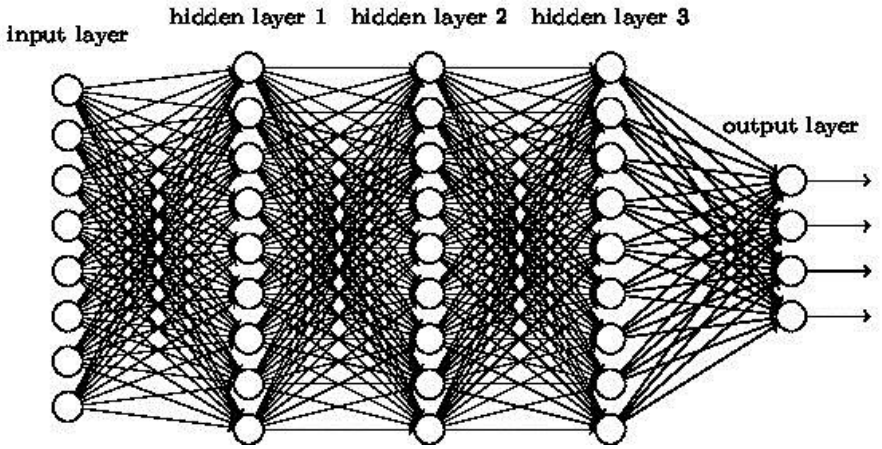
Walter Pitts

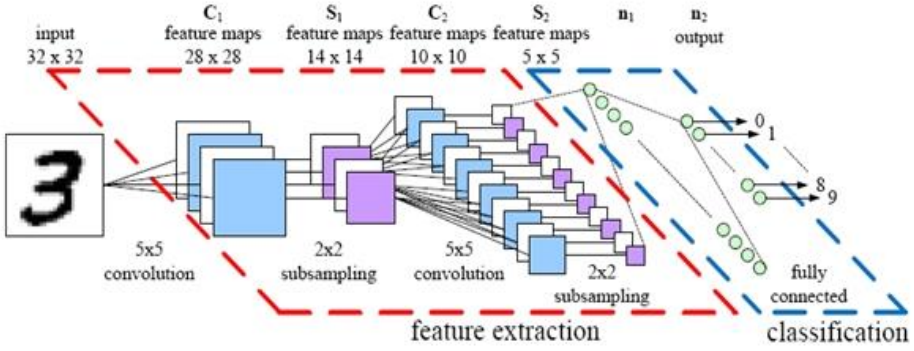


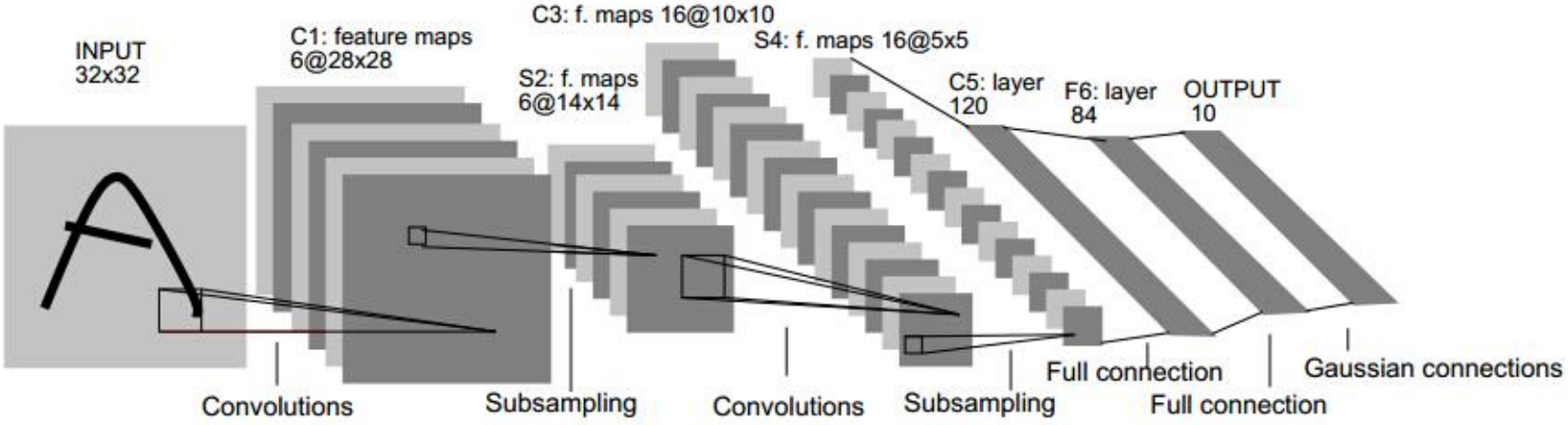
A cartoon drawing of a biological neuron (left) and its mathematical model (right).

Analysis in R^n

- Deep neural networks

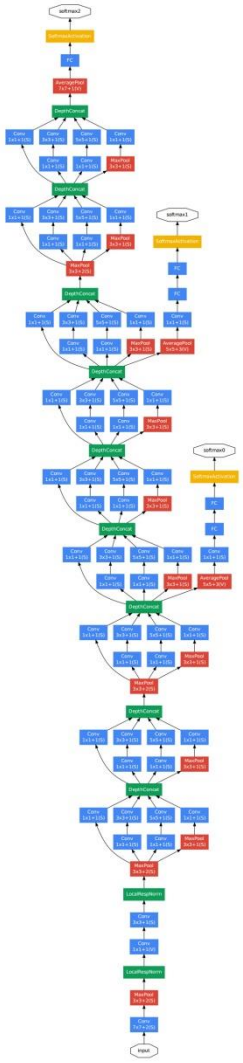


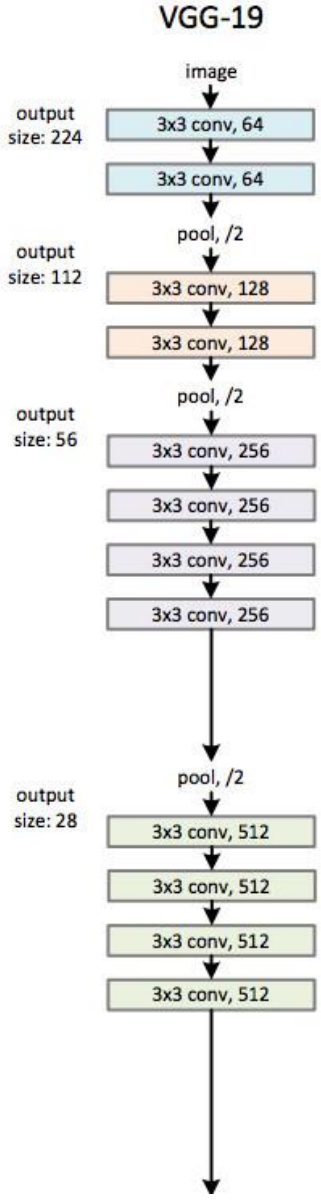


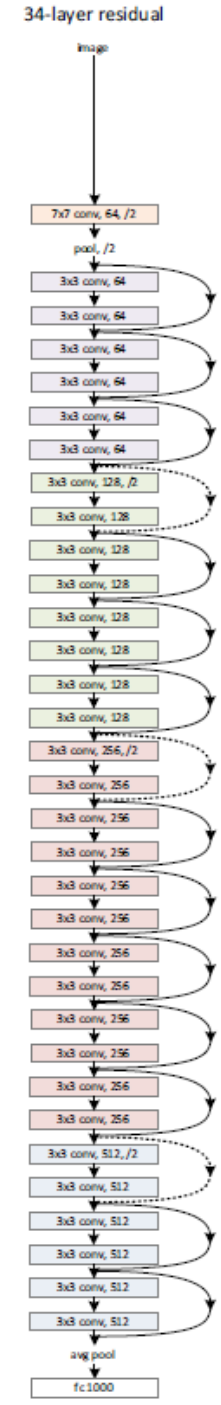


Analysis in R^n

- Deep neural networks







Analysis in R^n

$e(\mathbf{w}, \mathbf{v}, \mathbf{y}) = V(\mathbf{v}, \mathbf{y}, f(\mathbf{w}, \mathbf{v}))$

$$\frac{\partial e}{\partial \mathbf{w}} = \frac{\partial V}{\partial f}\frac{\partial f}{\partial \mathbf{w}} = \sum_{o \in \mathcal{O}} \frac{\partial V}{\partial f_o}\frac{\partial f_o}{\partial \mathbf{w}}.$$

$$\frac{\partial f_o}{\partial \mathbf{w}} = \sum_{o \in \mathcal{O}} \frac{\partial f_o}{\partial x_o}\frac{\partial x_o}{\partial \mathbf{w}}$$

- Backpropagation at each output node

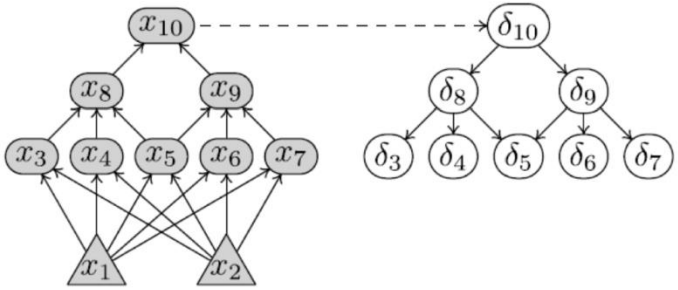
$w_{ij} : j \to i$

$$g_{ij}^o = \frac{\partial x_o}{\partial w_{ij}} = \frac{\partial x_o}{\partial a_i}\frac{\partial a_i}{\partial w_{ij}} = \frac{\partial x_o}{\partial a_i}\frac{\partial}{\partial w_{ij}} \sum_{h \in \mathrm{pa}(i)} w_{ih} x_h = \delta_i^o x_j, \qquad (2)$$

where $a_i = \sum_{h \in \mathrm{pa}(i)} w_{ih} x_h$ and we have defined $\delta_i^j := \partial x_j / \partial a_i$.

delta error

$$\delta_i^j = 0 \text{ whenever } i \succ j.$$



Output layer: $\delta_o^o = \sigma'(a_o)$, $\delta_j^o = 0$ if $j \notin \mathcal{O}$.

Hidden layer:

$$\delta_i^o = \frac{\partial x_o}{\partial a_i} = \sum_{h \in \mathrm{ch}(i)} \frac{\partial x_o}{\partial a_h}\frac{\partial a_h}{\partial x_i}\frac{\partial x_i}{\partial a_i} = \sigma'(a_i) \sum_{h \in \mathrm{ch}(i)} w_{hi}\delta_h^o. \qquad (3)$$

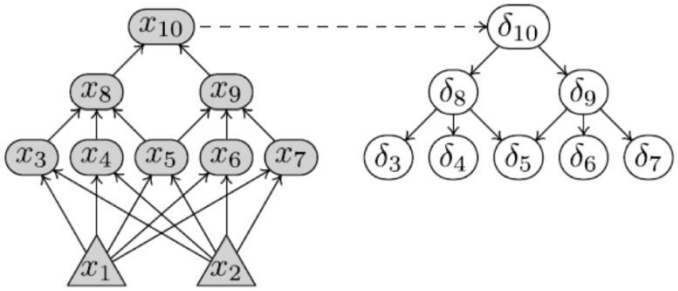
Analysis in R^n

- Backpropagation (simplified version)

$$\frac{\partial V}{\partial w_{ij}} = \frac{\partial V}{\partial a_i}\frac{\partial a_i}{\partial w_{ij}} = \delta_i x_j,$$

where this time δ_i is simply $\partial V / \partial a_i$.

$$\delta_i = \sum_{h \in \mathrm{ch}(i)} \frac{\partial V}{\partial a_h}\frac{\partial a_h}{\partial x_i}\frac{\partial x_i}{\partial a_i} = \sigma'(a_i) \sum_{h \in \mathrm{ch}(i)} w_{hi}\delta_h.$$



Analysis in R^n

- Backpropagation (simplified version)

Algorithm F (Forward propagation). FORWARD($\mathcal{G}, \mathbf{w}, \mathbf{m}, \mathbf{v}, \mathbf{x}$):

F1. [Initialize.] For all $i \in \mathcal{I}$ set $x_i \leftarrow v_i$ and initialize an integer variable $k \leftarrow 1$.

F2. [Topsort.] Invoke TOPSORT on $\mathcal{V} \backslash \mathcal{I}$, so that now the vector $\mathbf{s}$ contains the topological sorting of the nodes of the net. Set the variable l to the dimension of the vector $\mathbf{s}$.

F3. [Finished yet?] If $k \leq l$ go on to step F4, otherwise the algorithm stops.

F4. [Compute the state $\mathbf{x}$.] If $\mathbf{m} = (1, 1, \cdots, 1)^T$ set $x_{s_k} \leftarrow \sigma(\sum_{j \in \mathrm{pa}(s_k)} w_{s_k j} x_j)$ otherwise set $x_{s_k} \leftarrow m_{s_k} \sum_{j \in \mathrm{pa}(s_k)} w_{s_k j} x_j$. Increase k by one and go back to step F3.

for backpropagating errors using the same FORWARD subroutine

Analysis in R^n

- Backpropagation (simplified version)

Algorithm B (Backward propagation). BACKWARD($\mathcal{G}, \mathbf{w}, \mathbf{x}, q, V$):

B1. [Loss or output?] If $q \leq 0$ go to step B2, otherwise jump to step B3.

B2. [Initialize the loss.] For all $o \in \mathcal{O}$ set $v_o \leftarrow \partial V / \partial a_o$ and go to step B4.

B3. [Initialize x_q.] For each $o \in \mathcal{O}$ if $o \neq q$ set $v_o \leftarrow 0$, else if $o = q$ make the assignment $v_o \leftarrow \sigma'(\sigma^{-1}(x_o))$.

B4. [Compute backwards.] For each $k \in \mathcal{V} \backslash \mathcal{I}$ set $m_k \leftarrow \sigma'(\sigma^{-1}(x_k))$, then invoke FORWARD($(\mathcal{G}\backslash\mathcal{I})^T, \mathbf{w}^T, \mathbf{m}, \mathbf{v}, \boldsymbol{\delta}$).

inverting the direction of the graph and weight

B5. [Output the gradient.] For each $i \in \mathcal{V} \backslash \mathcal{I}$ and then for each $j \in \mathrm{pa}(i)$ set $g_{ij} \leftarrow \delta_i x_j$ and output g_{ij}. Terminate the algorithm.

Algorithm FB (Backpropagation).

FB1. [Forward] Invoke FORWARD($\mathcal{G}, \mathbf{w}, (1, 1, \cdots, 1)^T, \mathbf{v}, \mathbf{x}$).

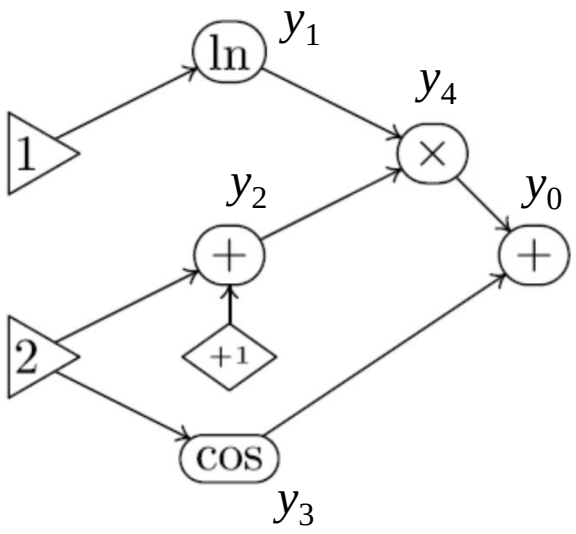
FB2. [Backward] Invoke BACKWARD($\mathcal{G}, \mathbf{w}, \mathbf{x}, -1, V$).

Analysis in R^n

- Automatic differentiation

$$y_o = f(x_1, x_2) = (1 + x_2) \ln x_1 + \cos x_2.$$

expression DAG



$$\begin{aligned}
&\text{(i)} && y_{0,1} = y_{3,1}\partial y_0/\partial y_3 + y_{4,1}\partial y_0/\partial y_4,\\
&\text{(ii)} && y_{3,1} = 0,\\
&\text{(iii)} && y_{4,1} = y_{1,1}\partial y_4/\partial y_1 + y_{2,1}\partial y_4/\partial y_2,\\
&\text{(iv)} && y_{1,1} = x_1^{-1},\\
&\text{(v)} && y_{2,1} = 0,\\
&\text{(vi)} && y_{0,2} = y_{3,2}\partial y_0/\partial y_3 + y_{4,2}\partial y_0/\partial y_4,\\
&\text{(vii)} && y_{3,2} = -\sin x_2,\\
&\text{(viii)} && y_{4,2} = y_{1,2}\partial y_4/\partial y_1 + y_{2,2}\partial y_4/\partial y_2,\\
&\text{(ix)} && y_{1,2} = 0,\\
&\text{(x)} && y_{2,2} = 1,
\end{aligned}$$

Two different differentiations: $y_{i,j} = \partial y_i/\partial x_j$, with $j = 1, 2$, and $\partial y_i/\partial y_k$.

$$\partial y_0/\partial y_3 = 1, \quad \partial y_0/\partial y_4 = 1, \quad \partial y_4/\partial y_1 = y_2, \quad \partial y_4/\partial y_2 = y_1.$$

$$y_{0,1} \rightsquigarrow \{\text{(iv)}, \text{(v)}, \text{(ii)}\}, \text{(iii)}, \text{(i)}, \quad y_{0,2} \rightsquigarrow \{\text{(vii)}, \text{(ix)}, \text{(x)}\}, \text{(viii)}, \text{(vi)}$$

Linear Algebra

- Norms
  - Inner product

The *standard inner product* on $\mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

In particular, the inner product between matrices is

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{tr}(\mathbf{X}^T \mathbf{Y}).$$

Linear Algebra

- Norms
  - Norms and distances

A function $f : \mathbb{R}^n \to \mathbb{R}$ with $\operatorname{dom} f = \mathbb{R}^n$ is called a *norm* if

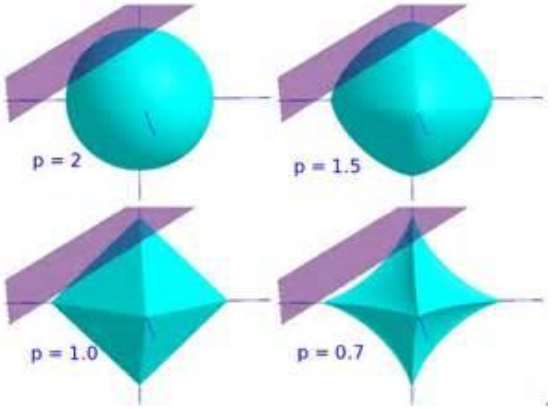
- f is nonnegative: $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$;
- f is definite: $f(\mathbf{x}) = 0$ only if $\mathbf{x} = \mathbf{0}$;
- f is homogeneous: $f(t\mathbf{x}) = |t|f(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$;
- f satisfies the triangle inequality: $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x})+f(\mathbf{y})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

We use the notation $f(\mathbf{x}) = \|\mathbf{x}\|$.

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

Examples: Euclidean norm, ℓ_p-norm, Mahalanobis norm, Frobenius norm, matrix 1-norm, matrix ∞-norm, matrix 2-norm

Linear Algebra



- Norms
  - Unit balls

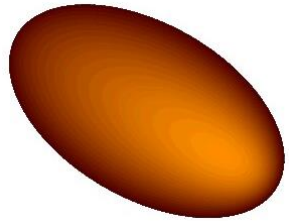
$$\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n | \, \|\mathbf{x}\| \leq 1\},$$

is called the *unit ball* of the norm $\|\cdot\|$. The unit ball satisfies the following properties:

- $\mathcal{B}$ is symmetric about the origin, *i.e.*, $\mathbf{x} \in \mathcal{B}$ if and only if $-\mathbf{x} \in \mathcal{B}$;
- $\mathcal{B}$ is convex;
- $\mathcal{B}$ is closed, bounded, and has nonempty interior.

Conversely, if $C \subseteq \mathbb{R}^n$ is any set satisfying these three conditions, then it is the unit ball of a norm, which is given by

$$\|\mathbf{x}\| = (\sup \{t \geq 0 | t\mathbf{x} \in C\})^{-1}.$$



Examples: unit balls of Euclidean norm, ℓ_p-norm, Mahalanobis norm

Linear Algebra

- Norms
  - Equivalence of norms

Suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$. A basic result of analysis is that there exist positive constants α and β such that, for all $\mathbf{x} \in \mathbb{R}^n$,

$$\alpha \|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq \beta \|\mathbf{x}\|_a .$$

This means that the norms are *equivalent*, *i.e.*, they define the same set of open subsets, the same set of convergent sequences, and so on

Examples: equivalence between Euclidean norm, ℓ_p-norm, and Mahalanobis norm

Linear Algebra

- Norms
  - Operator/induced norms

Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. We define the *operator norm* of $\mathbf{X} \in \mathbb{R}^{m \times n}$, induced by the norms $\|\cdot\|_a$ and $\|\cdot\|_b$, as

$$\|\mathbf{X}\|_{a,b} = \sup \{\|\mathbf{X}\mathbf{u}\|_a \mid \|\mathbf{u}\|_b \leq 1\}.$$

Examples: $\|\mathbf{X}\|_2$, $\|\mathbf{X}\|_\infty$, $\|\mathbf{X}\|_1$

Linear Algebra

- Norms
  - Nuclear norm & (p,q)-norm

The nuclear norm $\|\mathbf{A}\|_*$ of a matrix $\mathbf{A}$ is defined as the sum of all the singular values of $\mathbf{A}$. It is usually used for approximating the rank of a matrix.

Proposition 1. *Equivalence between matrix norms:*

$$1. \quad \|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \|\mathbf{A}\|_*.$$
$$2. \quad \|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1 \|\mathbf{A}\|_\infty}.$$

The (p, q)-norm of matrices is widely used in sparse representation:

$$\|\mathbf{A}\|_{p,q} = \left(\sum_{i=1}^{n} \|\mathbf{A}_i\|_p^q \right)^{\frac{1}{q}},$$

where $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_n)$ and $p, q \geq 1$.
Note that we use the same notation as the induced norm.

Linear Algebra

- Norms
  - Dual norm

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. The associated *dual norm*, denoted $\|\cdot\|_*$, is defined as

$$\|\mathbf{z}\|_* = \sup\{\mathbf{z}^\top \mathbf{x} | \, \|\mathbf{x}\| \leq 1\}.$$

Generalized Cauchy-Schwartz inequality:

$$\mathbf{z}^\top \mathbf{x} \leq \|\mathbf{x}\| \, \|\mathbf{z}\|_* \, .$$

$$\|\mathbf{x}\|_{**} = \|\mathbf{x}\|$$

Examples: Euclidean norm, ℓ_p norm, matrix 2-norm.

Linear Algebra

- Symmetric eigenvalue decomposition (EVD)

Suppose $\mathbf{A} \in \mathbb{S}^n$, *i.e.*, $\mathbf{A}$ is a real symmetric $n \times n$ matrix. Then $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top, \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is *orthogonal*, *i.e.*, satisfies $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. The (real) numbers λ_i are the *eigenvalues* of $\mathbf{A}$, and are the roots of the *characteristic polynomial* $\det(s\mathbf{I} - \mathbf{A})$. The columns of $\mathbf{Q}$ form an orthonormal set of *eigenvectors* of $\mathbf{A}$. The factorization (1) is called the *spectral decomposition* or (symmetric) *eigenvalue decomposition* of $\mathbf{A}$.

We order the eigenvalues as $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_n$. We use the notation $\lambda_i(\mathbf{A})$ to refer to the ith largest eigenvalue of $\mathbf{A} \in \mathbb{S}^n$. We usually write the largest or maximum eigenvalue as $\lambda_1(\mathbf{A}) = \lambda_{\max}(\mathbf{A})$, and the least or minimum eigenvalue as $\lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$.

$$\det \mathbf{A} = \prod_{i=1}^n \lambda_i, \ \ \mathrm{tr}\, \mathbf{A} = \sum_{i=1}^n \lambda_i, \ \ \|\mathbf{A}\|_2 = \max\{\lambda_1, -\lambda_n\}, \ \ \|\mathbf{A}\|_F = \left(\sum_{i=1}^n \lambda_i^2\right)^{1/2}.$$

Linear Algebra

- Definiteness

$$\lambda_{\max}(\mathbf{A}) = \sup_{x \neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}, \quad \lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{x} \neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}.$$

A matrix $\mathbf{A} \in \mathbb{S}^n$ is called *positive definite* if for all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$. We denote this as $\mathbf{A} \succ \mathbf{0}$. By the inequality above, we see that $\mathbf{A} \succ \mathbf{0}$ if and only all its eigenvalues are positive, *i.e.*, $\lambda_{\min}(\mathbf{A}) > 0$. If $-\mathbf{A}$ is positive definite, we say $\mathbf{A}$ is *negative definite*, which we write as $\mathbf{A} \prec \mathbf{0}$. We use $\mathbb{S}^n_{++}$ to denote the set of positive definite matrices in $\mathbb{S}^n$.
If $\mathbf{A}$ satisfies $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$, we say that $\mathbf{A}$ is *positive semidefinite* or *nonnegative definite*. If $-\mathbf{A}$ is in nonnegative definite, *i.e.*, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq 0$ for all $\mathbf{x}$, we say that $\mathbf{A}$ is *negative semidefinite* or *nonpositive definite*. We use $\mathbb{S}^n_{+}$ to denote the set of nonnegative definite matrices in $\mathbb{S}^n$.
For $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$, we use $\mathbf{A} \prec \mathbf{B}$ to mean $\mathbf{B} - \mathbf{A} \succ \mathbf{0}$, and so on.

Linear Algebra

- Symmetric squareroot

Let $\mathbf{A} \in \mathbb{S}_+^n$, with eigenvalue decomposition $\mathbf{A} = \mathbf{Q}\,\mathrm{diag}(\lambda_1, ..., \lambda_n)\mathbf{Q}^\top$. We define the (symmetric) squareroot of $\mathbf{A}$ as

$$\mathbf{A}^{1/2} = \mathbf{Q}\,\mathrm{diag}(\lambda_1^{1/2}, ..., \lambda_n^{1/2})\mathbf{Q}^\top.$$

The squareroot $\mathbf{A}^{1/2}$ is the unique symmetric positive semidefinite solution of the equation $\mathbf{X}^2 = \mathbf{A}$.

Linear Algebra

- Singular value decomposition (SVD)

Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $\operatorname{rank} \mathbf{A} = r$. Then $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{m \times r}$ satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, $\mathbf{V} \in \mathbb{R}^{n \times r}$ satisfies $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, ..., \sigma_r)$, with

$$\sigma_{\max}(\mathbf{A}) = \sup_{\mathbf{x},\mathbf{y} \neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \sup_{\mathbf{y} \neq 0} \frac{\|\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{y}\|_2}.$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The factorization (1) is called the *singular value decomposition* (SVD) of $\mathbf{A}$. The columns of $\mathbf{U}$ are called *left singular vectors* of $\mathbf{A}$, the columns of $\mathbf{V}$ are *right singular vectors*, and the numbers σ_i are the *singular values*. The singular value decomposition can be written

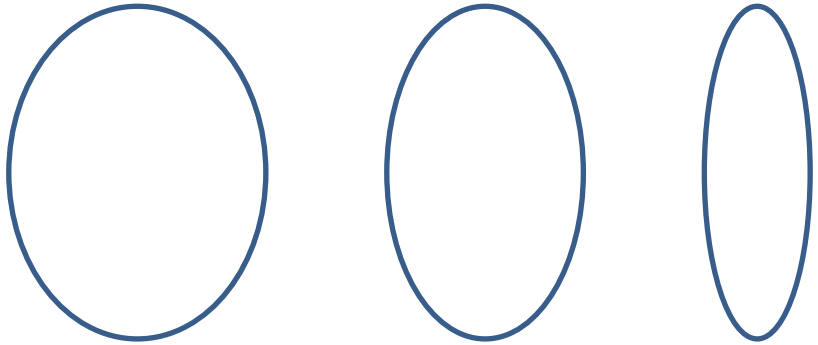
$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Linear Algebra

- Condition number

The *condition number* of a nonsingular $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $\text{cond}(\mathbf{A})$ or $\kappa(\mathbf{A})$, is defined as

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2 = \sigma_{\max}(\mathbf{A})/\sigma_{\min}(\mathbf{A}).$$



Linear Algebra

- Pseudo-inverse

Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{A} \in \mathbb{R}^{m\times n}$, with $\operatorname{rank}\mathbf{A} = r$. We define the *pseudo-inverse* or *Moore-Penrose inverse* of $\mathbf{A}$ as

$$\mathbf{A}^\dagger = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top \in \mathbb{R}^{n\times m}.$$

Linear Algebra

- Adjoint operator

Given a linear operator $\mathcal{A} : \mathbb{R}^m \to \mathbb{R}^n$, its *adjoint operator* is defined as the linear operator $\mathcal{A}^*$ that satisfies:

$$\langle \mathcal{A}^*(\mathbf{x}), \mathbf{y} \rangle = \langle \mathbf{x}, \mathcal{A}(\mathbf{y}) \rangle, \quad \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m.$$

Examples: $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, $\mathcal{A}(\mathbf{X}) = \mathbf{x}$ is the linear operator that extracts entries from $\mathbf{X}$.

Linear Algebra

- von Neumann trace theorem

Theorem 1. *Suppose* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$. *Then*

$$|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \sum_{i=1}^{\min(m,n)} \sigma_i(\mathbf{A}) \sigma_i(\mathbf{B}).$$

In particular, when both $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ p.s.d. matrices, the inequality becomes

$$\sum_{i=1}^{n} \sigma_i(\mathbf{A}) \sigma_{n-i+1}(\mathbf{B}) \leq \operatorname{tr}(\mathbf{AB}) \leq \sum_{i=1}^{n} \sigma_i(\mathbf{A}) \sigma_i(\mathbf{B}).$$

Homework (2)

1. For each of the following sequences, determine the rate of convergence and the rate constant.

   a. $x_k = 2^{-k}$, for $k = 1, 2, \cdots$.

   b. $x_k = 1 + 5 \times 10^{-2k}$, for $k = 1, 2, \cdots$.

   c. $x_k = 2^{-2^k}$.

   d. $x_k = 3^{-k^2}$.

   e. $x_k = 1 - 2^{-2^k}$ for k odd, and $x = 1 + 2^{-k}$ for k even.

2. Let $\{x_k\}$ and $\{c_k\}$ be convergent sequences, and assume that

$$\lim_{k \to \infty} c_k = c \neq 0.$$

Consider the sequence $\{y_k\}$ with $y_k = c_k x_k$. Can its convergence rate and rate constant be determined from those of $\{x_k\}$ and $\{c_k\}$?

Homework (2)

3. Compute the gradient and Hessian of the following functions:

   a. $f(\mathbf{x}) = \|\mathbf{x}\|_p$, $\mathbf{x} \neq \mathbf{0}$, $p > 1$.

   b. $f(\mathbf{x}) = (\mathbf{a}^T\mathbf{x})(\mathbf{b}^T\mathbf{x})$.

   c. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

Homework (2)

4. Compute the gradient of the following matrix functions (write in matrices, rather than entrywise. Give details. Working out 5 is sufficient if you don't want to torture yourself.):

a. $f(\mathbf{X}) = \|\mathbf{X}\|_F$.

b. $f(\mathbf{X}) = \|\mathbf{X}^T\mathbf{A}\mathbf{X}\|_F^2$ and $\mathbf{A}$ is a symmetric matrix.

c. $f(\mathbf{X}) = \|\operatorname{diag}(\mathbf{X}^T\mathbf{A}\mathbf{X})\|_F^2$ and $\mathbf{A}$ is a symmetric matrix.

d. $f(\mathbf{X}) = \det \mathbf{X}$.

e. $f(\mathbf{X}) = \det(\mathbf{X}^T\mathbf{A}\mathbf{X})$.

f. $f(\mathbf{X}) = \mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}$.

g. $f(\mathbf{X}) = \operatorname{tr}(\mathbf{A}\mathbf{X}\mathbf{B})$.

h. $f(\mathbf{X}) = \operatorname{tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})$.

i. $f(\mathbf{X}) = \operatorname{tr}((\mathbf{A}+\mathbf{X})^{-1})$.

Homework (2)

5. Write a program for computing the gradient of error function V with respect to the weights of a neural network with arbitrary topology. Test it on simplified ResNet: 1 filter, 1 skip connection, no pooling and batch normalization. Verify the correctness by numerical differentiation, i.e. verify

$$\frac{V(\mathbf{w}+t\mathbf{v})-V(\mathbf{w})}{t} \approx \langle \nabla V(\mathbf{w}), \mathbf{v} \rangle$$

for arbitrary choice of $\mathbf{v}$ and sufficiently small t. The code should include the verification step.

6. Write an automatic differentiation program that works on a given expression DAG. Test it on function y_o and verify it with numerical differentiation:

$$y_o = (\sin(x_1{+}1){+}\cos(2x_2))\tan(\log(x_3)){+}(\sin(x_2{+}1){+}\cos(2x_1))\exp(1{+}\sin(x_3)).$$

The code should include the verification step.

Homework (2)

7. Find the dual norm of Mahalanobis norm: $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^T\mathbf{M}\mathbf{x}}$, where $\mathbf{M}$ is a positive definite matrix.
8. Prove that the eigenvalues λ_i of $(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}$, where $\mathbf{A}$ is positive semidefinite and $\mathbf{B}$ is positive definite, satisfy $0 \leq \lambda_i < 1$.
9. Compute the condition number of the following matrix:

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 6 & 9 \end{bmatrix}.$$

10. Suppose $\mathbf{X} \in \mathbb{R}^{3\times 3}$, $\mathcal{A}(\mathbf{X}) = X_{11} + X_{12} - X_{31} + 2X_{33}$, find $\mathcal{A}^*$.