CSCE 790: Neural Networks and Their Applications

AIISC and Dept. Computer Science and Engineering
Email: vignar@sc.edu

September 12, 2023

UNIVERSITY OF
South Carolina

Recap

- ML - Functional view of models
- Models - Parametric models
- Multi-layer feedforward neural network
- Learning paradigm - Supervised learning
    - Classification
    - Regression

Math Recap

- Set and operations on set (e.g., Cartesian product)
- Relation $\rightarrow$ Functions
- Vector space $(V, F, +, .)$
- Span
- Linear independence and Basis
- Inner product
- Norm

Example: Regression

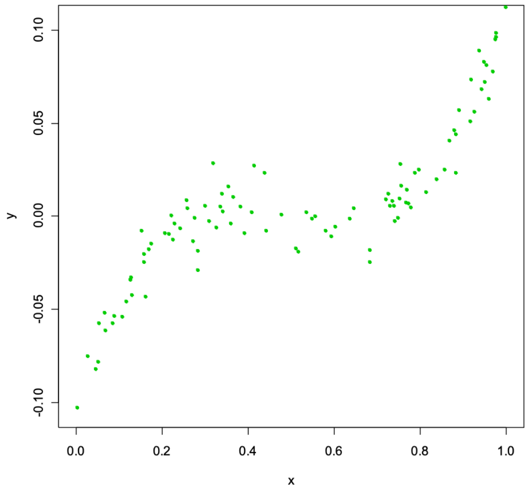


Figure: Input-output data points

- $\{(x, y)\}$ - Given set of input-output pairs
- $x \in \Omega \subseteq \mathbb{R}$ and $y \in \mathbb{R}$
- Prediction function $f : \Omega \to \mathbb{R}$
- Input space $(\Omega, \mathbb{R}, +, .)$
- Through ML we are searching for a function in the function space(?) - a set of (continuous?) functions from $\Omega \to \mathbb{R}$

Example: Regression

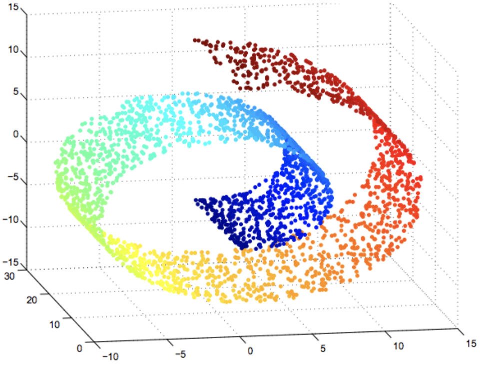


Figure: Input-output data points in a 3-D space

- $\{(x, y)\}$ - Given set of input-output pairs
- $x \in \Omega \subseteq \mathbb{R}^2$ and $y \in \mathbb{R}$
- Prediction function $f : \Omega \to \mathbb{R}$
- Input space $(\Omega, \mathbb{R}, +, .)$
- Here we are searching for a function in the function space - a set of continuous functions from $\Omega \to \mathbb{R}$

Math Recap

- Linear vs Nonlinear functions
- Optimization problems - (Cost function, Decision variables, and Constraint set)
- Taylor's formula and its relevance

Recap

- Taylor's expansion provides a representation of function as infinite sum of terms expressed as the functions derivatives at a single point
- Mean value theorem provides a way to stop the expansion after the first derivative
- Theorem of Extended value of mean provides a way to stop the expansion after the second derivative
- If $f''(\theta) > 0, \forall\, \theta$, and $f'(\theta^*) = 0$,

$$\implies f(\theta) = f(\theta^*) + 0 + \text{a positive number} \quad \forall\, \theta \neq \theta^*$$
$$\implies f(\theta) > f(\theta^*) \quad \forall\, \theta \neq \theta^*$$
$$\implies \theta^* \text{ is the minimizer of } f(\theta)$$

Minimizers

- Minimizers
- Local vs Global minimizers

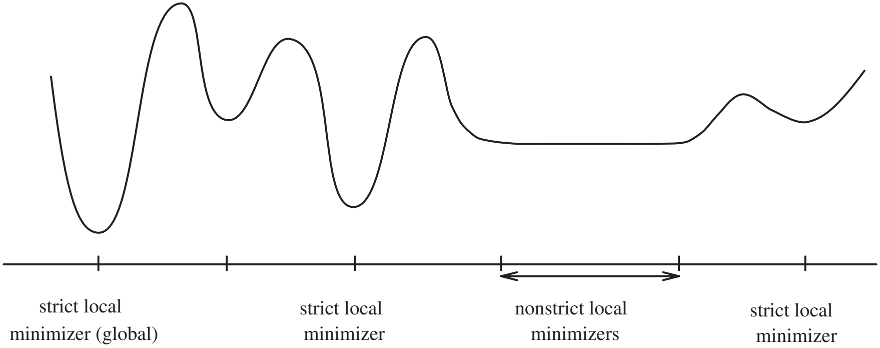


Figure: Examples of local minimizers

Math Recap - Moving beyond one dimension

- Symmetric matrices
- Positive definiteness
- Eigen values and spectral radius
- Notion of 'local' or 'neighborhood' (To be defined)

Derivatives

Definition (Derivative)

Let $\Theta \subset \mathbb{R}$ and let $f : \Theta \to \mathbb{R}$ be a real-valued function. Suppose that Θ contains a neighborhood of the point θ. We define the derivative of f at θ by

$$f'(\theta) = \lim_{\alpha \to 0} \frac{f(\theta + \alpha) - f(\theta)}{\alpha}$$

provided that the limit exists. In that case we say that f is differentiable at θ.

Visualization (1-D)

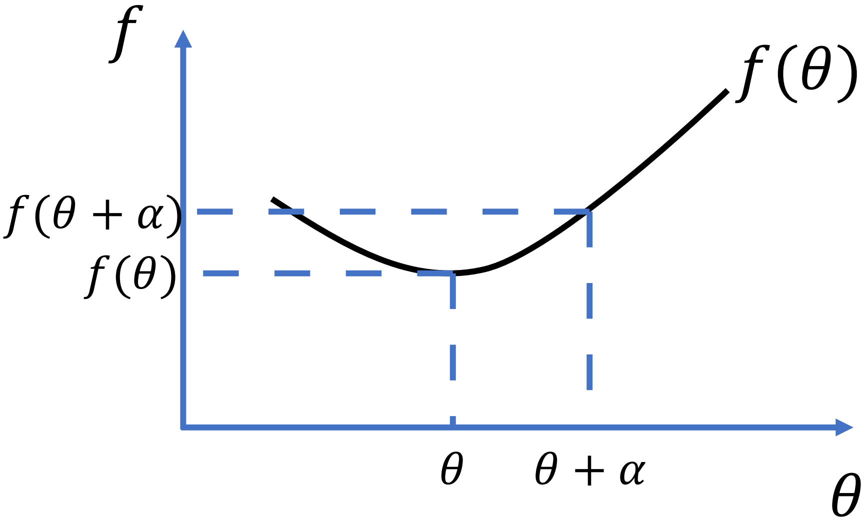


Figure: Changes in domain and range - Derivatives in 1-D

Derivatives

The definition above does not work when we pass from functions of single real variable to functions of several real variables. Now when $\Theta \subset \mathbb{R}^n$ and $f : \Theta \to \mathbb{R}^m$, we have

$$\frac{\overbrace{f(\theta+\alpha) - f(\theta)}^{\in\mathbb{R}^m}}{\underbrace{\alpha}_{\in\mathbb{R}^n}}.$$

We do not know what it means to divide by a vector and hence should seek another definition. Modify the definition of a derivative to accommodate vector-valued functions of several real variables.

Derivatives

Definition (Directional derivative)

Let $\Theta \subset \mathbb{R}^n$ and $f : \Theta \to \mathbb{R}^m$. Suppose that Θ contains a neighborhood of θ. Given $d \in \mathbb{R}^n$ with $d \neq 0$, define

$$f'(\theta; d) = \lim_{\alpha \to 0} \frac{f(\theta + \alpha d) - f(\theta)}{\alpha},$$

provided the limit exists. It is called the directional derivative of f at θ with respect to d.

Derivatives

Definition (Partial derivative)

If

$$f'(\theta; e_i) = \lim_{\alpha \to 0} \frac{f(\theta + \alpha e_i) - f(\theta)}{\alpha}$$

exists it is called the i^{th} partial derivative of f at θ, denoted $\dfrac{\partial f(\theta)}{\partial \theta_i}$.

Definition (Gradient)

Assume that $\dfrac{\partial f(\theta)}{\partial \theta_i}$ exists $\forall i$. The gradient of f at θ is defined as

$$\nabla f(\theta) = \begin{bmatrix} \frac{\partial f(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial f(\theta)}{\partial \theta_n} \end{bmatrix} \qquad \text{and } \nabla f(\theta) = (Df(\theta))'.$$

Derivatives

Definition (Hessian)

Suppose that $\dfrac{\partial f(\theta)}{\partial \theta_i} \in C^\infty$ is a continuously differentiable function of θ. The Hessian matrix of f at $\theta \in \Theta$ is given by

$$H(\theta) = \nabla^2 f(\theta) = \begin{bmatrix} \dfrac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_1} & \dfrac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_n} \\ \dfrac{\partial^2 f(\theta)}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 f(\theta)}{\partial \theta_2 \partial \theta_2} & \cdots & \dfrac{\partial^2 f(\theta)}{\partial \theta_2 \partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\theta)}{\partial \theta_n \partial \theta_1} & \dfrac{\partial^2 f(\theta)}{\partial \theta_n \partial \theta_2} & \cdots & \dfrac{\partial^2 f(\theta)}{\partial \theta_n \partial \theta_n} \end{bmatrix} \in S^n,$$

since $\frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 f(\theta)}{\partial \theta_j \partial \theta_i}$ for $i, j = 1, \ldots, n$.

Example

Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(\theta_1, \theta_2) = \theta_1^3 - 12\theta_1\theta_2 + 8\theta_2^3$. Let $\theta = (\theta_1, \theta_2)$, and then

$$\nabla f(\theta) = \begin{bmatrix} \frac{\partial f(\theta)}{\partial \theta_1} \\ \frac{\partial f(\theta)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 3\theta_1^2 - 12\theta_2 \\ -12\theta_1 + 24\theta_2^2 \end{bmatrix}$$

and

$$H(\theta) = \begin{bmatrix} 6\theta_1 & -12 \\ -12 & 48\theta_2 \end{bmatrix}.$$

Derivatives

Definition (Gradient matrix)

If $\Theta \subset \mathbb{R}^n$ and $f : \Theta \to \mathbb{R}^m$ is a vector-valued function, i.e., $f(\theta) = (f_1(\theta), \ldots, f_m(\theta))'$, then f is called differentiable if f_i is differentiable for all $i = 1, \ldots, m$. The gradient matrix of f is the $n \times m$ matrix

$$\nabla f(\theta) = \left[\begin{array}{c|c|c} \nabla f_1(\theta) & \cdots & \nabla f_m(\theta) \end{array} \right]_{n \times m} = (J(\theta))'$$

where $J(\theta)$ is the Jacobian of f.

Derivatives

Example

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(\theta_1, \theta_2) = \theta_1 \theta_2$. The directional derivatives of f at $a = (a_1, a_2)$ with respect to

1. $d_1 = (1, 0)'$ is
$$f'(a; d_1) = \lim_{\alpha \to 0} \frac{(a_1 + \alpha)a_2 - a_1 a_2}{\alpha} = a_2.$$
2. $d_2 = (1, 2)'$ is
$$f'(a; d_2) = \lim_{\alpha \to 0} \frac{(a_1 + \alpha)(a_2 + 2\alpha) - a_1 a_2}{\alpha} = 2a_1 + a_2.$$

Derivatives

Definition (Open ball)

For all norms $\|\cdot\|$ in $\mathbb{R}^n$ and for any $\varepsilon > 0$, we define an open ball or ε-neighborhood of $\theta_0 \in \mathbb{R}^n$ by $B_\varepsilon(\theta_0) = B(\theta_0, \varepsilon) = \{\theta \in \mathbb{R}^n : \|\theta - \theta_0\| < \varepsilon\}$.

Example

The unit ball $B(0, 1)$ in $\mathbb{R}^2$ contains all the points inside a circle of radius one centered at the origin.

Derivatives

Theorem (Extended M.V.T: Taylor's theorem, second order expansions)

Let $B(\theta, r)$ be an open ball centered at $\theta \in \mathbb{R}^n$ with radius r. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable (C^2) over $B(\theta, r)$. Then

1. *For all y such that $\theta + y \in B(\theta, r)$, i.e., $\|y\| < r$, there exists an $\alpha \in [0, 1]$ such that*

$$f(\theta + y) = f(\theta) + y'\nabla f(\theta) + \frac{1}{2} y' \nabla^2 f(\theta + \alpha y) y.$$

2. *For all y such that $\theta + y \in B(\theta, r)$ there holds*

$$f(\theta + y) = f(\theta) + y'\nabla f(\theta) + \frac{1}{2} y' \nabla^2 f(\theta) y + o(\|y\|^2).$$

Local vs. Global minima

Definition

Let $\mathcal{F} = \{\theta \in \mathbb{R}^n \mid g_i(\theta) \leq 0, h_j(\theta) = 0, i = 1, \ldots, m, j = 1, \ldots, \ell\}$ be the feasible region of a NLP.

1. $\theta^* \in \mathcal{F} \subset \mathbb{R}^n$ is a local minimum of the NLP if there exists $\varepsilon > 0$ such that $f(\theta^*) \leq f(\theta)$ for any $\theta \in B(\theta, \varepsilon) \cap \mathcal{F}$.
2. $\theta^* \in \mathcal{F}$ is a strict local minimum of the NLP if there exists $\varepsilon > 0$ such that $f(\theta^*) < f(\theta)$ for any $\theta \in B(\theta, \varepsilon) \cap \mathcal{F}$, $\theta \neq \theta^*$.
3. $\theta^* \in \mathcal{F}$ is a global minimum of the NLP if $f(\theta^*) \leq f(y)$ for $\forall y \in \mathcal{F}$.
4. $\theta^* \in \mathcal{F}$ is a strict global minimum of the NLP if $f(\theta^*) < f(y)$ for $\forall y \in \mathcal{F}$, $y \neq \theta^*$.

Unconstrained Optimization

Now, we examine algorithms for unconstrained optimization, which are motivated by moving from a point θ along a descent direction d with step size $\alpha > 0$ and repeating until $\nabla f(\theta^*) = 0$. A first order approximation can be used. The central idea is based on Taylor's expansion

$$f(\theta + \alpha d) \approx f(\theta) + \alpha(\nabla f(\theta))' d,$$

and if $(\nabla f(\theta))' d < 0$, then $f(\theta + \alpha d) < f(\theta)$ for some $\alpha > 0$. Let's start with an interesting and fundamental observation of descent directions.

Proposition (Descent directions)

Let $f : \mathbb{R}^n \to \mathbb{R}$ *be differentiable at* θ*. If there exists a* $d \in \mathbb{R}^n$ *such that* $(\nabla f(\theta))' d < 0$*, then* $\forall \alpha > 0$ *sufficiently small,* $f(\theta + \alpha d) < f(\theta)$*. We call* d *the descent direction and* α *the step size.*

Unconstrained Optimization

Definition (Level Set)

A level set of a real-valued function f of n variables is a set of the form

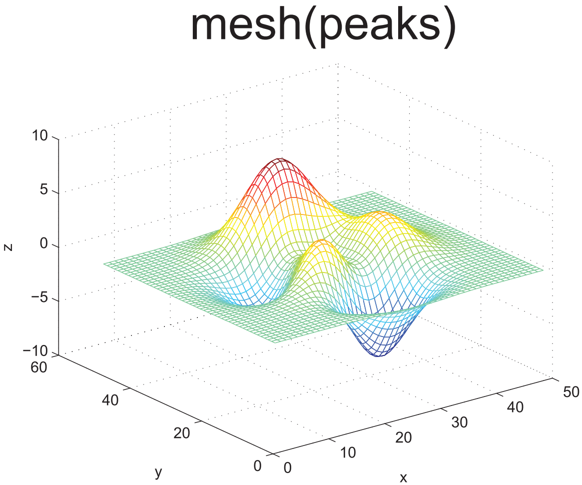
$$L_c(f) = \{(\theta_1, \ldots, \theta_n)' \mid f(\theta_1, \ldots, \theta_n) = c\}.$$

Note, conventionally, associated with a convex function f one can define a level set, sometimes called a *lower-level set*,

$$S_\alpha = \{\theta \in S \mid f(\theta) \leq \alpha\}, \quad \alpha \in \mathbb{R},$$

to differentiate it from the *upper-level set* $\{\theta \in S \mid f(\theta) \geq \alpha\}$.

Unconstrained Optimization



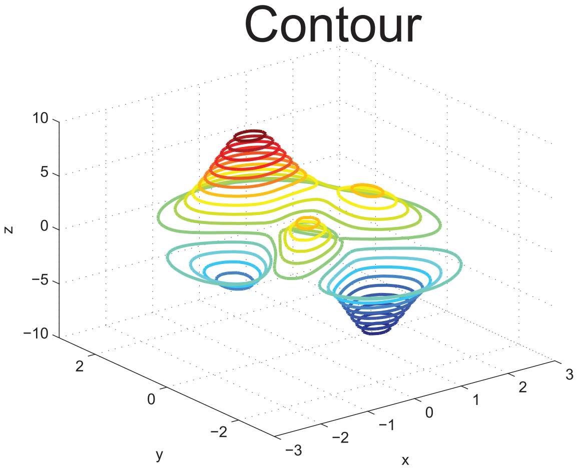


Figure: The level sets of "Peaks"

Unconstrained Optimization

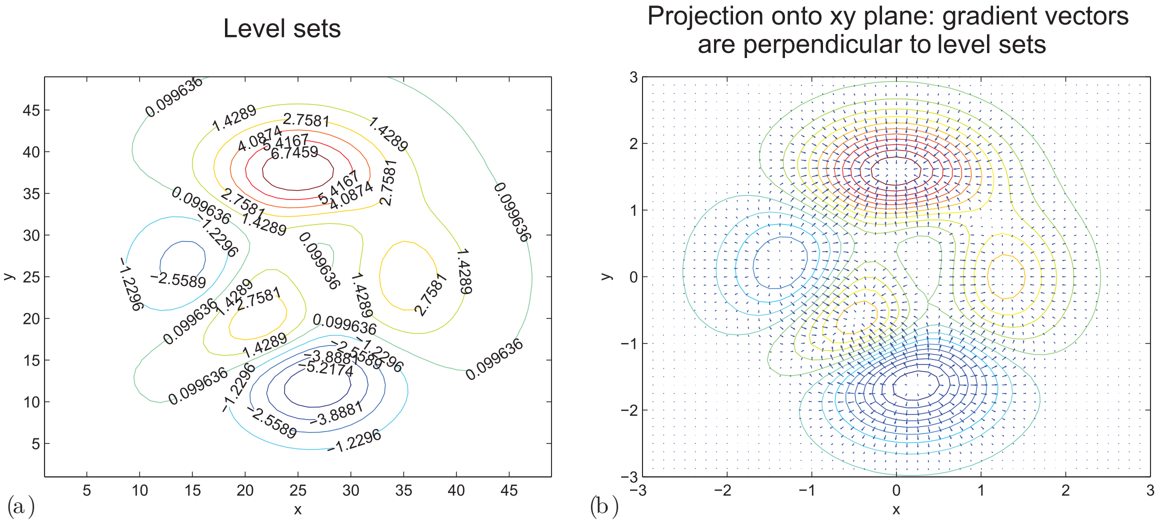


Figure: Gradients are perpendicular to the level sets

Unconstrained Optimization

In the Figure,

$$\begin{aligned} z = f(x,y) =& 3(1-x)^2 e^{-x^2-(y+1)^2} \\ & - 10\left(\frac{x}{5} - x^3 - y^5\right) e^{-x^2-y^2} - \frac{1}{3} e^{-(x+1)^2-y^2}, \end{aligned}$$

is a function of two variables, obtained by translating and scaling Gaussian distributions.

Gradient Methods

1. **Motivation:** Decrease $f(\theta)$ until $\nabla f(\theta^*) = 0$ based on

$$f(\theta + \alpha d) \approx f(\theta) + \alpha(\nabla f(\theta))' d.$$

   If $(\nabla f(\theta))' d < 0$, then $f(\theta + \alpha d) < f(\theta)$ for small $\alpha > 0$.

2. **Procedure:** We start at some point θ^0 (an **initial guess**) and successively generate vectors θ^1, θ^2, ..., such that f is decreased at each iteration, that is, $f(\theta^{k+1}) < f(\theta^k)$ for all $k = 0, 1, 2, \ldots$.

Algorithms for Unconstrained Optimization

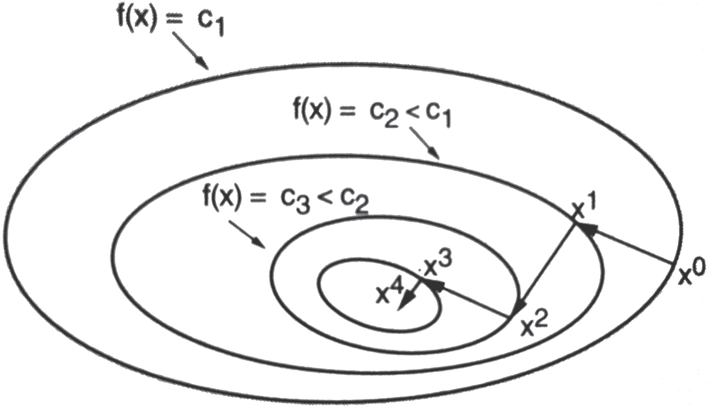


Figure: Iterative Descent.

Gradient-based Iterative Algorithms (Generic)

Proposition (Gradient is orthogonal to level set of a function)

The gradient of f at a point is perpendicular to the level set of f at that point.

Remark:

Therefore, if the direction d makes an angle with $\nabla f(\theta)$ that is greater than 90°, that is,

$$(\nabla f(\theta))' d < 0,$$

there is an interval $(0, \delta)$ of step sizes such that

- $f(\theta + \alpha d) < f(\theta), \quad \forall \alpha \in (0, \delta),$
- $$\cos(\theta) = \frac{(\nabla f(\theta))' \cdot d}{\|\nabla f(\theta)\| \cdot \|d\|} < 0 \implies \theta > 90^\circ.$$

Gradient-based Iterative Algorithms (Generic)

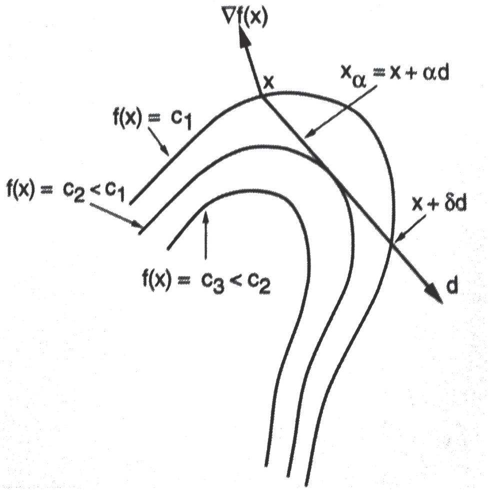


Figure: Orthogonality of Gradient to Level Sets.

Gradient-based Iterative Algorithms (Generic)

Algorithm (Generic algorithm)

At each iteration k,

- $\theta^{k+1} = \theta^k + \alpha^k d^k$
- *If $\nabla f(\theta^k) \neq 0$, then the direction d^k is chosen so that $(\nabla f(\theta^k))' d^k < 0$.*
- *The step size $\alpha^k > 0$ is chosen such that $f(\theta^k + \alpha^k d^k) < f(\theta^k)$.*
- *Principal example:*

$$\theta^{k+1} = \theta^k - \alpha^k D^k \nabla f(\theta^k), \quad d^k = -D^k \nabla f(\theta^k),$$

  - $D^k \succ 0$.
  - $(\nabla f(\theta^k))' \cdot d^k = (\nabla f(\theta^k))'(-D^k \nabla f(\theta^k)) < 0$.

Neural network weight selection and training

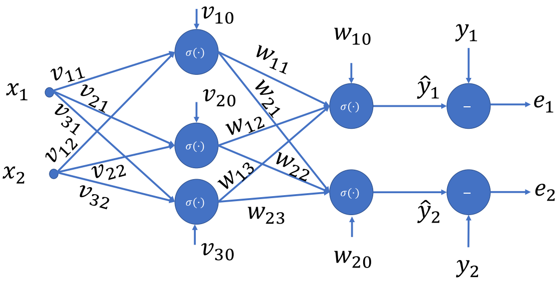


Figure: Error-credit assignment problem

- For a NN to function as desired, their weights and biases need to be selected appropriately
- It was for many years unknown, how to use the error to tune the weights of each layer - 'error-credit assignment problem'

Lemma: Chain rule

Proposition

Let $f : \mathbb{R}^k \to \mathbb{R}^m$ *and* $g : \mathbb{R}^m \to \mathbb{R}^n$ *be smooth, i.e.,* C^∞*. Let* $h : \mathbb{R}^k \to \mathbb{R}^n$ *be defined by* $h(\theta) = g(f(\theta))$*. Then*

$$\nabla h(\theta) = \nabla f(\theta) \nabla g(f(\theta)), \quad \forall \theta \in \mathbb{R}^k.$$