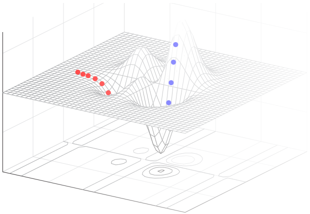
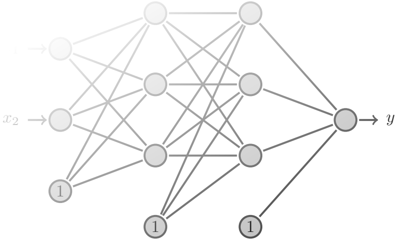
MLP trainieren

Prof. Dr. Jörg Frochte

Maschinelles Lernen





Loss-Funktion

- Folgende Aspekte bilden die Architektur eines neuronalen Netzwerkes:
  - Anzahl der Schichten
  - Anzahl der Neuronen pro Schicht
  - Auswahl der Aktivierungsfunktionen
- Sind sie einmal gewählt, sind diese Merkmale nicht Teil des Trainings.
- Das Training versucht die Gewichte durch eine numerische Optimierung zu bestimmen.
- Man sucht das Minimum der Loss-Funktion

$$J(\mathbf{W}) = \frac{1}{|D|} \sum_{(\mathbf{x}_d, y_d) \in D} \left(y_d - \hat{y}(\mathbf{x}_d, \mathbf{W})\right)^2,$$

mit $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots\}$

- Der Ansatz, wie ein Minimum von J in Abhängigkeit von W gefunden werden soll, läuft bei uns im Folgenden über das Gradientenverfahren.
- Der **Gradient** einer Funktion f kann mithilfe des sogenannten Nabla-Operators

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}$$

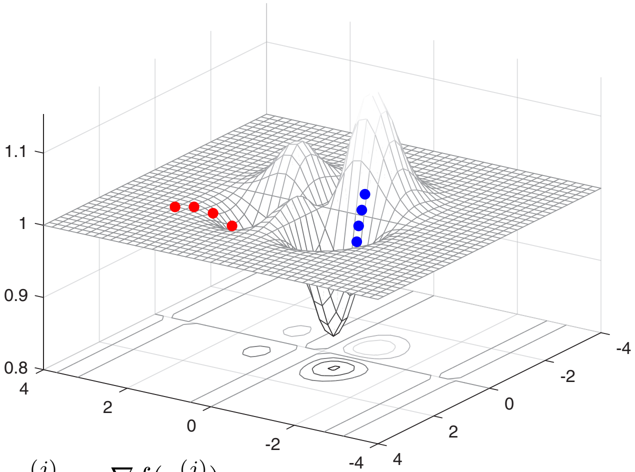
ausgedrückt werden. Dabei ist $\frac{\partial}{\partial x_1}$ als Operation eine **partielle Ableitung**.

Beispiel

Wenn $f(x,y) = 3x^2 - y^2 + xy$ ist, lautet der Gradient:

$$\nabla f = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 6x + y \\ -2y + x \end{pmatrix}$$

- Beim Gradientenverfahren macht man es sich zunutze, dass der Gradient ∇f einer Funktion $f : \mathbb{R}^n \to \mathbb{R}$ in Richtung des steilsten Anstiegs zeigt.
- Entsprechend zeigt $-\nabla f$ in Richtung des steilsten Abstiegs.
- Um in Richtung des steilsten Abstiegs zu gehen, wird beim Gradientenverfahren wie folgt iteriert:



$$x^{(j+1)} = x^{(j)} - \eta \nabla f(x^{(j)})$$

- Man startet bei einem Punkt $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$, der kein Extremum sein sollte.
- Betrachtet man die Abbildung oben erkennt man zwei Probleme.

$$x^{(j+1)} = x^{(j)} - \eta \nabla f(x^{(j)})$$

- η ist dabei ein Parameter zwischen $]0,1]$, der die Stabilität des Verfahrens erhöhen kann, wodurch jedoch höhere Kosten entstehen.

Pseudocode des Gradientenverfahren

Require: $x^{(1)}$, m

1: $k = 1$
2: $\eta = 0.1$
3: **while** $\nabla f(x^{(k)}) \neq 0$ **AND** $k < m$ **do**
4: $\quad x^{(k+1)} = x^{(k)} - \eta \nabla f(x^{(k)})$
5: $\quad k = k + 1$
6: **end while**

- Die Abbruchbedingung in Zeile 3 ist natürlich theoretisch.
- In der Praxis benutzt man $\|\nabla f(x^{(k)})\| \leq \varepsilon$.

- Wir werden dieses Verfahren jetzt auf unsere Fehlerfunktion unten anwenden.

$$J(W) = \sum_{(x_d, y_d) \in D} \frac{1}{2} \left(y_d - y(x_d, W)\right)^2$$

- Das führt zu folgender Regel für das Update der Gewichte

$$W_{neu} = W_{alt} - \sigma \nabla J(W_{alt})$$

- Wir müssen also den Gradienten unserer Fehlerfunktion bzgl. W bestimmen: $\nabla J(W)$

$$\nabla J(W) = \nabla \sum_{(x_d, y_d) \in D} \frac{1}{2} \left(y_d - y(x_d, W)\right)^2$$

- Da wir eine gegebene Menge an Beispielen haben und diese nicht von W abhängen, ist y_d für uns in jedem dieser Summanden eine Konstante.

$$
\begin{aligned}
\nabla J(W) &= \nabla \sum_{(x_d, y_d) \in D} \frac{1}{2} \left(y_d - y(x_d, W)\right)^2 \\
&= \sum_{(x_d, y_d) \in D} \underbrace{\left(y_d - y(x_d, W)\right)}_{\text{äußere Ableitung}} \cdot \underbrace{\left(-\nabla y(x_d, W)\right)}_{\text{innere Ableitung}}
\end{aligned}
$$

- Da wir eine gegebene Menge an Beispielen haben und diese nicht von W abhängen, ist y_d für uns in jedem dieser Summanden eine Konstante.
- Bei Umformung tritt die bekannte Kettenregel auf mit der inneren Ableitung $-\nabla y$.

$$\begin{aligned}\nabla J(W) &= \nabla \sum_{(x_d,y_d)\in D} \frac{1}{2}\left(y_d - y(x_d,W)\right)^2 \\ &= \sum_{(x_d,y_d)\in D} \underbrace{(y_d - y(x_d,W))}_{\text{äußere Ableitung}} \cdot \underbrace{(-\nabla y(x_d,W))}_{\text{innere Ableitung}} \\ &= -\sum_{(x_d,y_d)\in D} (y_d - y(x_d,W))\, \nabla y(x_d,W)\end{aligned}$$

- Da wir eine gegebene Menge an Beispielen haben und diese nicht von W abhängen, ist y_d für uns in jedem dieser Summanden eine Konstante.
- Bei Umformung tritt die bekannte Kettenregel auf mit der inneren Ableitung $-\nabla y$.

$$
\begin{aligned}
\nabla J(W) &= \nabla \sum_{(x_d, y_d) \in D} \frac{1}{2} \left(y_d - y(x_d, W) \right)^2 \\
&= \sum_{(x_d, y_d) \in D} \left(y_d - y(x_d, W) \right) \cdot \left(-\nabla y(x_d, W) \right) \\
&= -\sum_{(x_d, y_d) \in D} \left(y_d - y(x_d, W) \right) \nabla y(x_d, W)
\end{aligned}
$$

- Da wir eine gegebene Menge an Beispielen haben und diese nicht von W abhängen, ist y_d für uns in jedem dieser Summanden eine Konstante.
- Bei Umformung tritt die bekannte Kettenregel auf mit der inneren Ableitung $-\nabla y$.
- Nun gilt es den Gradienten von y bzgl. W zu berechnen.
- Wir schreiben zur besseren Übersicht im Folgenden auch y anstatt $y(x_d, W)$ und analog mit den Neuronenausgängen $O_i^{(j)}$.

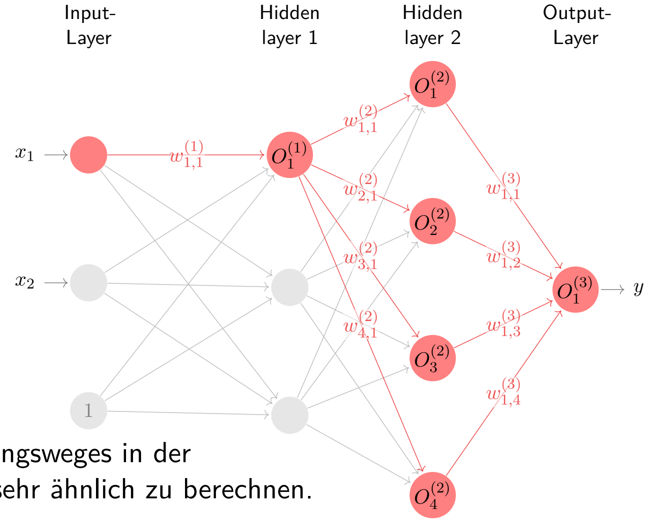
Von $w_{1,1}^{(i)}$ beeinflusster Ausschnitt des Netzes

Einige der nötigen partiellen Ableitungen sind:

$$\frac{\partial J(W)}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(3)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(1)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(1)}}$$



Die Größen, nach denen hier differenziert wird, liegen entlang des obersten Verbindungsweges in der Abbildung. Die anderen Ableitungen sind sehr ähnlich zu berechnen.

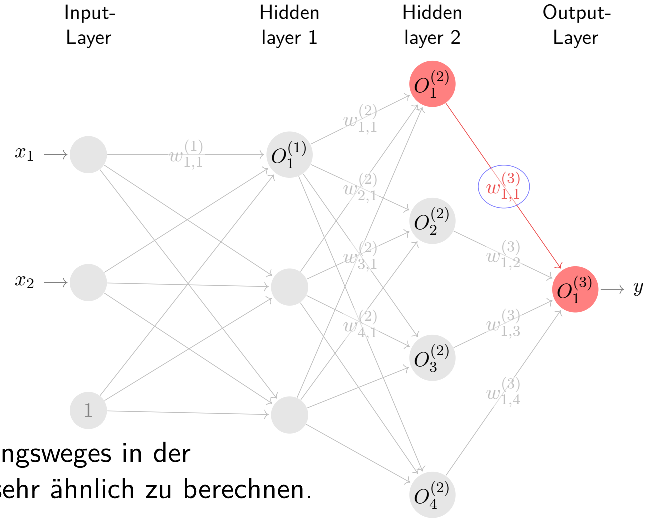
Von $w_{1,1}^{(i)}$ beeinflusster Ausschnitt des Netzes

Einige der nötigen partiellen Ableitungen sind:

$$\frac{\partial J(W)}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(3)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(1)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(1)}}$$



Die Größen, nach denen hier differenziert wird, liegen entlang des obersten Verbindungsweges in der Abbildung. Die anderen Ableitungen sind sehr ähnlich zu berechnen.

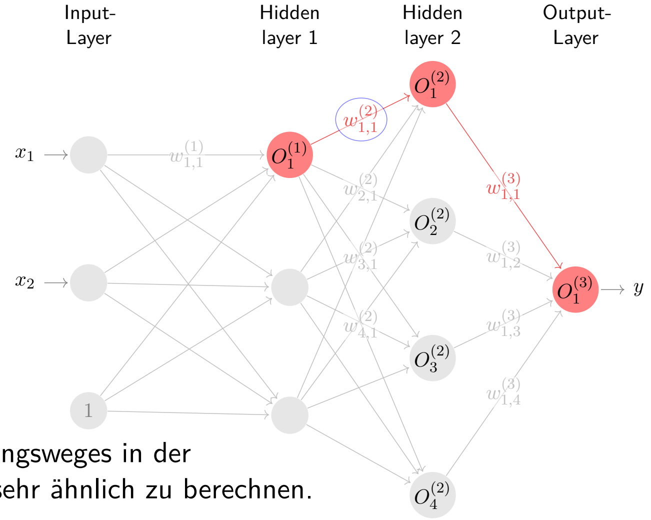
Von $w_{1,1}^{(i)}$ beeinflusster Ausschnitt des Netzes

Einige der nötigen partiellen Ableitungen sind:

$$\frac{\partial J(W)}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(3)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(1)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(1)}}$$



Die Größen, nach denen hier differenziert wird, liegen entlang des obersten Verbindungsweges in der Abbildung. Die anderen Ableitungen sind sehr ähnlich zu berechnen.

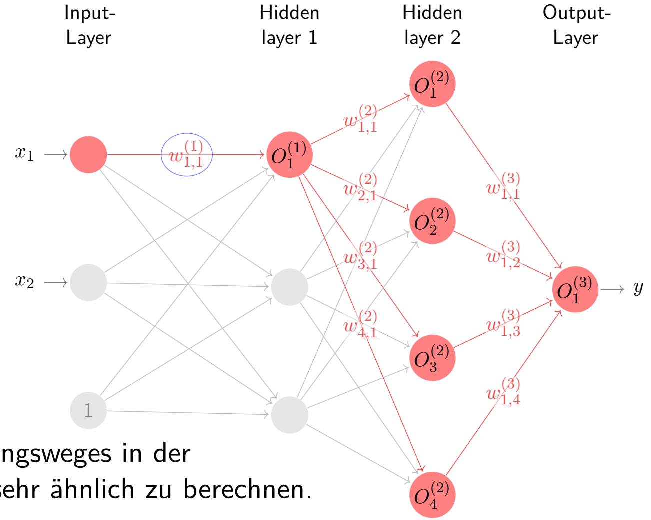
Von $w_{1,1}^{(i)}$ beeinflusster Ausschnitt des Netzes

Einige der nötigen partiellen Ableitungen sind:

$$\frac{\partial J(W)}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(3)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(1)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(1)}}$$



Die Größen, nach denen hier differenziert wird, liegen entlang des obersten Verbindungsweges in der Abbildung. Die anderen Ableitungen sind sehr ähnlich zu berechnen.

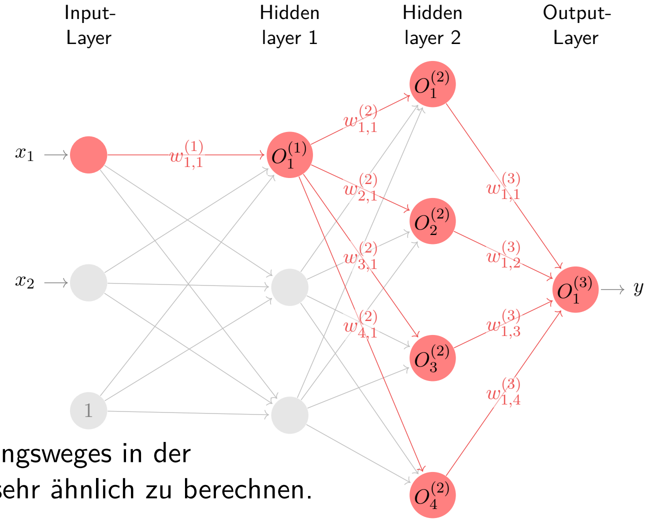
Von $w_{1,1}^{(i)}$ beeinflusster Ausschnitt des Netzes

Einige der nötigen partiellen Ableitungen sind:

$$\frac{\partial J(W)}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(3)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}}$$

$$\frac{\partial J(W)}{\partial w_{1,1}^{(1)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(1)}}$$



Die Größen, nach denen hier differenziert wird, liegen entlang des obersten Verbindungsweges in der Abbildung. Die anderen Ableitungen sind sehr ähnlich zu berechnen.

Ableitung der Ausgangsschicht

- Es ist am einfachsten, die am weitesten rechts stehende Schicht zu betrachten, und sich dann nach links durchzuarbeiten.
- Wir gehen nun entsprechend vor:

$$y = O_1^{(3)} = w_{1,1}^{(3)} \cdot O_1^{(2)} + w_{1,2}^{(3)} \cdot O_2^{(2)} + w_{1,3}^{(3)} \cdot O_3^{(2)} + w_{1,4}^{(3)} \cdot O_4^{(2)}$$

$$\Longrightarrow \frac{\partial y}{\partial w_{1,1}^{(3)}} = O_1^{(2)}$$

$$\Longrightarrow \frac{\partial J(W)}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(3)}} = -\sum_D (y_d - y) \cdot O_1^{(2)}$$

- Die weiteren Ableitungen in dieser Schicht sind analog zu bilden:

$$\frac{\partial y}{\partial w_{1,j}^{(3)}} = O_j^{(2)} \qquad \Longrightarrow \qquad \frac{\partial J(W)}{\partial w_{1,j}^{(3)}} = -\sum_D (y_d - y) \cdot O_j^{(2)}$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)} \right) \cdot O_1^{(1)} \right]$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)} \right) \cdot O_1^{(1)} \right]$$

- Nebenrechnung (1/3):

$$\frac{\partial y}{\partial O_1^{(2)}} = \frac{\partial \left(w_{1,1}^{(3)} \cdot O_1^{(2)} + w_{1,2}^{(3)} \cdot O_2^{(2)} + w_{1,3}^{(3)} \cdot O_3^{(2)} + w_{1,4}^{(3)} \cdot O_4^{(2)} \right)}{\partial O_1^{(2)}} = w_{1,1}^{(3)}$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[\underbrace{O_1^{(2)} \left(1 - O_1^{(2)}\right)}_{\text{äußere Ableitung}} \cdot O_1^{(1)} \right]$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)} \right) \cdot \underbrace{O_1^{(1)}}_{\text{innere Ableitung}} \right]$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)}\right) \cdot O_1^{(1)} \right]$$

- Nebenrechnung (2/3): Die äußere Ableitung ergibt sich durch die Ableitung der Sigmoid-Funktion. Es gilt: $\mathrm{sig}'(x) = \mathrm{sig}(x)(1 - \mathrm{sig}(x))$.

$$O_1^{(2)} = \mathrm{sig}(\dots + O_1^{(1)} \cdot w_{1,1}^{(2)} + \dots), \qquad u(v) = \mathrm{sig}(v), \quad v(W) = \dots + O_1^{(1)} w_{1,1}^{(2)} + \dots$$

$$\frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = \underbrace{\mathrm{sig}(\dots + O_1^{(1)} w_{1,1}^{(2)} + \dots)}_{=O_1^{(2)}} \Big(1 - \underbrace{\mathrm{sig}(\dots + O_1^{(1)} w_{1,1}^{(2)} + \dots)}_{=O_1^{(2)}}\Big) \cdot \frac{\partial v}{\partial w_{1,1}^{(2)}}$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)} \right) \cdot O_1^{(1)} \right]$$

- Nebenrechnung (3/3): Die innere Ableitung von $O_1^{(2)}$ nach $w_{1,1}^{(2)}$ ist

$$\frac{\partial v}{\partial w_{1,1}^{(2)}} = \frac{\partial (\cdots + O_1^{(1)} w_{1,1}^{(2)} + \dots)}{\partial w_{1,1}^{(2)}} = O_1^{(1)}$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)}\right) \cdot O_1^{(1)} \right]$$

$$\Longrightarrow \frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y)\, w_{1,1}^{(3)}\, O_1^{(2)} \left(1 - O_1^{(2)}\right) O_1^{(1)}$$

Ableitung der vorletzten Schicht

- In den weiteren Schichten müssen wir Gebrauch von der Kettenregel machen.
- Erinnerung: Sei $u(v(x))$ eine zusammengesetzte Funktion dann ist $\frac{\partial u}{\partial x} = \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial x}$
- So erhalten wir in der nächsten Schicht, wobei $w_{1,1}^{(3)}$ hier die Ableitung $\frac{\partial y}{\partial O_1^{(2)}}$ ist.

$$\frac{\partial y}{\partial w_{1,1}^{(2)}} = \frac{\partial y}{\partial O_1^{(2)}} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(2)}} = w_{1,1}^{(3)} \cdot \left[O_1^{(2)} \left(1 - O_1^{(2)}\right) \cdot O_1^{(1)} \right]$$

$$\Longrightarrow \frac{\partial J(W)}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y) \cdot \frac{\partial y}{\partial w_{1,1}^{(2)}} = -\sum_D (y_d - y)\, w_{1,1}^{(3)}\, O_1^{(2)} \left(1 - O_1^{(2)}\right) O_1^{(1)}$$

- Allgemein sind die partiellen Ableitungen dieser Schicht:

$$\frac{\partial y}{\partial w_{j,k}^{(2)}} = w_{1,j}^{(3)} \cdot O_j^{(2)} \left(1 - O_j^{(2)}\right) \cdot O_k^{(1)} \Rightarrow \frac{\partial J(W)}{\partial w_{j,k}^{(2)}} = -\sum_D (y_d - y)\, w_{1,j}^{(3)}\, O_j^{(2)} \left(1 - O_j^{(2)}\right) O_k^{(1)}$$

Ableitung der vorderen Schicht

- Wir notieren y entsprechend der letzten Schicht und arbeiten uns nach vorne.

$$\frac{\partial y}{\partial w_{1,1}^{(1)}} = \frac{\partial}{\partial w_{1,1}^{(1)}} \left(w_{1,1}^{(3)} \cdot O_1^{(2)} + w_{1,2}^{(3)} \cdot O_2^{(2)} + w_{1,3}^{(3)} \cdot O_3^{(2)} + w_{1,4}^{(3)} \cdot O_4^{(2)} \right)$$

$$= w_{1,1}^{(3)} \frac{\partial O_1^{(2)}}{\partial w_{1,1}^{(1)}} + w_{1,2}^{(3)} \frac{\partial O_2^{(2)}}{\partial w_{1,1}^{(1)}} + w_{1,3}^{(3)} \frac{\partial O_3^{(2)}}{\partial w_{1,1}^{(1)}} + w_{1,4}^{(3)} \frac{\partial O_4^{(2)}}{\partial w_{1,1}^{(1)}} = \ldots$$

- Analog zum Ansatz zuvor wenden wir wieder die Kettenregel an, um uns in die nächste Schicht zurück zu begeben.
- Wie wir in der Abbildung gesehen haben, beeinflusst unser Gewicht $w_{1,1}^{(1)}$ die nachfolgende erste Schicht nur im Knoten $O_1^{(1)}$.

$$\cdots = w_{1,1}^{(3)} \frac{\partial O_1^{(2)}}{\partial O_1^{(1)}} \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} + w_{1,2}^{(3)} \frac{\partial O_2^{(2)}}{\partial O_1^{(1)}} \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} + w_{1,3}^{(3)} \frac{\partial O_3^{(2)}}{\partial O_1^{(1)}} \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} + w_{1,4}^{(3)} \frac{\partial O_4^{(2)}}{\partial O_1^{(1)}} \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} = \ldots$$

- Nun sollten wir dazu übergehen, das mit Summenzeichen zu schreiben um eine allgemeine Form für alle Gewichte gewinnen zu können.

$$\cdots = \sum_j w_{1,j}^{(3)} \frac{\partial O_j^{(2)}}{\partial O_1^{(1)}} \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} = \ldots$$

- Nun sollten wir dazu übergehen, das mit Summenzeichen zu schreiben um eine allgemeine Form für alle Gewichte gewinnen zu können.

$$\dots = \sum_j w_{1,j}^{(3)} \frac{\partial O_j^{(2)}}{\partial O_1^{(1)}} \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} = \dots$$

- Der letzte Faktor hängt nicht von j ab. Wir können die Ableitung also direkt bilden und dann vorziehen. Dabei nutzen wir aus, dass Folgendes gilt:

$$O_1^{(1)} = \mathrm{sig}(W^{(1)}x) \Rightarrow \frac{\partial O_1^{(1)}}{\partial w_{1,1}^{(1)}} = \mathrm{sig}(W^{(1)}x)\left(1 - \mathrm{sig}(W^{(1)}x)\right)\underbrace{x_1}_{(*)} = O_1^{(1)}\left(1 - O_1^{(1)}\right)x_1$$

- Der Term $(*)$ ist die beim Differenzieren auftretende innere Ableitung. Nun nutzen wir das, um unsere angefangene Rechnung fortzusetzen:

$$\dots = \sum_j w_{1,j}^{(3)} \frac{\partial O_j^{(2)}}{\partial O_1^{(1)}} O_1^{(1)}\left(1 - O_1^{(1)}\right)x_1 = O_1^{(1)}(1 - O_1^{(1)})x_1 \sum_j w_{1,j}^{(3)} \frac{\partial O_j^{(2)}}{\partial O_1^{(1)}} = \dots$$

$$\ldots = O_1^{(1)}\left(1 - O_1^{(1)}\right) x_1 \sum_j w_{1,j}^{(3)} \frac{\partial O_j^{(2)}}{\partial O_1^{(1)}}$$

$$= O_1^{(1)}\left(1 - O_1^{(1)}\right) x_1 \sum_j w_{1,j}^{(3)} \frac{\partial \operatorname{sig}(\cdots + O_1^{(1)} \cdot w_{j,1}^{(2)} + \ldots)}{\partial O_1^{(1)}} = \ldots$$

- Für den letzten Term bilden wir jetzt wieder die Ableitung und können dabei auch ausnutzen, dass es sich wieder um eine Sigmoid-Funktion handelt.

$$\frac{\partial y}{\partial w_{1,1}^{(1)}} = \cdots = O_1^{(1)}\left(1 - O_1^{(1)}\right) x_1 \sum_j w_{1,j}^{(3)} O_j^{(2)}\left(1 - O_j^{(2)}\right) w_{j,1}^{(2)}$$

- Allgemein sind die partiellen Ableitungen dieser Schicht:

$$\frac{\partial J(W)}{\partial w_{k,l}^{(1)}} = -\sum_D (y_d - y) \frac{\partial y}{\partial w_{k,l}^{(1)}} = -\sum_D (y_d - y)\, x_{d,l}\, O_k^{(1)}\left(1 - O_k^{(1)}\right) \sum_j w_{1,j}^{(3)} O_j^{(2)}\left(1 - O_j^{(2)}\right) w_{j,k}^{(2)}$$

- Mit diesen Ableitungen können wir nun die Ableitung der Fehlerfunktion berechnen:

$$\frac{\partial E(W)}{\partial W} = -\sum_{D} (y_d - y) \cdot \frac{\partial y}{\partial W} \tag{1}$$

- Es bleibt die Frage, was die Menge D hier konkret sein soll.
- Hierzu muss man sich den Unterschied zwischen **Batch-Learning** und **Incremental Learning** vor Augen führen.
- Beide Varianten beginnen damit, dass zunächst die Gewichte W zufällig initialisiert werden.
- Charakteristisch für das *Batch-Learning* ist, dass die Ableitung über eine Summe wie oben in Gleichung (1) erbracht wird. D – Teilmenge der Trainingsmenge – ist dabei der Batch. Das D kann dabei auch die ganze Trainingsmenge sein.
- Beim *incremental learning* wird immer nur ein Beispiel bzw. Datensatz betrachtet.
- Was sind die Vor- und Nachteile der beiden Ansätze?

- Wenn das inkrementelle Lernen zusammen mit dem Gradientenabstiegsverfahren verwendet wird nennt man es auch **Stochastic Gradient Descent** (SDG).
- Theoretisch scheint es so zu sein, dass nur das Batch-Learning über der ganzen Trainingsmenge einen vergleichsweise kontinuierlichen Abstieg sicherstellen kann.
- In der Praxis ist es so, dass sich besonders zu Beginn des Lernens die unterschiedlichen Fehler oft aufheben und es nur zu wenigen Fortschritten kommt.
- Die Gefahr ist groß, dass große Batches langsamer konvergieren und sich öfter in Nebenminima verirren.
- Anderseits ist reines *Incremental Learning* anfällig für Instabilitäten.
- Ein guter Ansatz ist daher oft, eher kleine Batches aus der größeren Trainingsmenge zu ziehen und mit diesen zu lernen.
- Leider beeinflusst die Größe des Batches auch die Parameterwahl vieler Verfahren.
- Wir gehen hier diesbezüglich nicht in die Tiefe und konzentrieren uns für die Eigenimplementierung auf den **Stochastic Gradient Descent** und nutzen für Batch-Ansätze Keras.

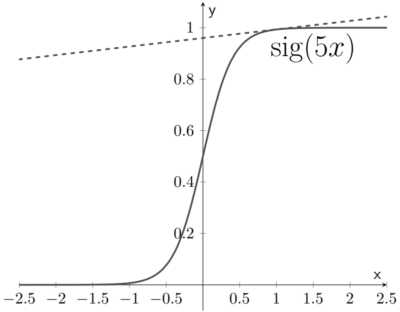
- Eng verbunden mit dem inkrementellen und dem Batch-Lernen sind die Begriffen des **Offline- und Online-Learnings**.
- Beim **Offline-Learning** werden alle Daten gespeichert und können jederzeit abgerufen werden.
- Beim **Online-Learning** wird jeder Datensatz nach der Bearbeitung verworfen und die Gewichte werden aktualisiert.
- Wir erhalten folgende Zusammenhänge:

$$\text{Batch-Lernen} \Rightarrow \text{Offline-Learning}$$
$$\text{Online-Learning} \Rightarrow \text{inkrementell Lernen}$$

Beispiele/Aufgabe

Denken Sie über Beispiele für ein Offline-Learning nach, das gleichzeitig inkrementell ist.

- Wir konzentrieren uns nun auf Probleme beim Training und vernachlässigen Aspekte bzgl. der Zusammenstellung von Batches.
- Ein Allgemeines Problem ist das der **Sättigung**.
- Große Aktivierungswerte – welche sich ja als Summe der Eingangssignale multipliziert mit den Gewichten ergeben – führen zu Gradienten mit sehr geringen Steigungen.
- Die Ableitung der Sigmoid-Funktion oder des Tangens Hyperbolicus für große Werte ist beinahe parallel zur x-Achse.
- In dem Beispiel hat die Ableitung bei $x = 1$ noch ungefähr den Wert 0.0332.



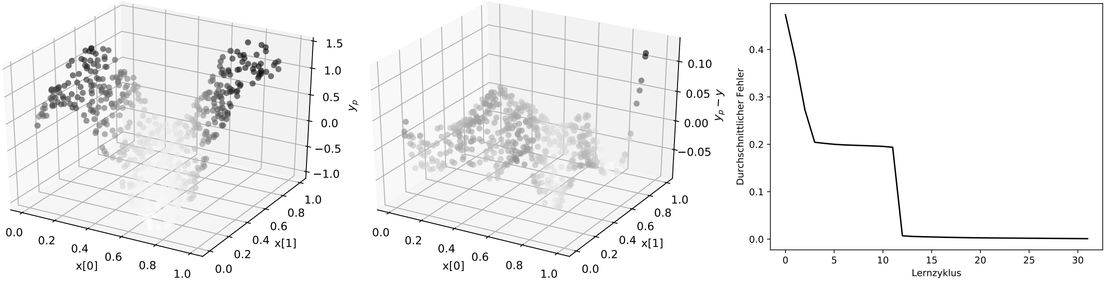
Gradient bzw. Tangente mit geringer Steigung bei $x = 1$

- Der Eintrag der Ableitung im Gradientenabstiegsverfahren verschwindet fast.
- Gehen also Neuronen in die Sättigung macht man es den Netzwerk schwer, die Gewichte zu ändern und so etwas *zu lernen*.

- Wir betrachten einmal das Training eines Netzes bzgl. auf der Funktion

$$y = \sin(2\pi(x + 0.5y)) + 0.5y$$

basierenden, verrauschen Trainingsdaten. Das Rauschen führt zu einem relativen Fehler von maximal 5% führen.



- In der Mitte die Differenz zwischen f und der Prognose. Rechts der Verlauf des Fehlers während des Trainings.
- Man sieht, dass Phasen, in denen länger nichts passiert, abrupt übergehen können in Phasen mit plötzlicher Verbesserung.