Agenda:

+ Multivariate Ableitungen (verschiedene Arten)
+ Kettenregel und der Anwendung in KI.
- Höhere Ableitungen
- Multivariate Optimierung

Wir haben eine Abbildung $f: \mathbb{R}^n \to \mathbb{R}^m$ (Der Definitions-bereich kann auch eine Teilmenge von $\mathbb{R}^n$ sein; Einfachheit halber hier $\mathbb{R}^n$).

$$f(x) = f(x_1, \dots, x_n) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ f_2(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{bmatrix}$$

Fall $m = 1$: $f: \mathbb{R}^n \to \mathbb{R}$ Funktion in n Variablen $x_1, \dots, x_n$

Fall $m = n$: Eine (nichtlineare) Transformation des Raums $\mathbb{R}^n$.

Die Jacobi-Matrix, bezeichnet als $J_f(x)$ oder $f'(x)$ oder $\dfrac{\partial(f_1, f_2, \dots, f_m)}{\partial(x_1, x_2, \dots, x_n)}$:

$$\begin{array}{c|c} & \begin{matrix} \frac{\partial}{\partial x_1} & \cdots & \frac{\partial}{\partial x_n} \end{matrix} \\ \begin{matrix} f_1 \\ \vdots \\ f_m \end{matrix} & \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \end{array}$$

Die Absicht: Die kleine Änderung von f,

$\Delta f := f(x+\Delta x) - f(x)$ zu approximieren

als $f'(x) \cdot \Delta x$.

Δx ist der Vektor der Änderungen der Variablen

$$\Delta x = \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix}.$$

Unter welcher Bedingung hat man die Approximation?

Es reicht aus, dass alle partiellen Ableitungen

$$\frac{\partial f_i}{\partial x_j} : \mathbb{R}^n \to \mathbb{R} \quad \text{stetig sind} \quad (i = 1, \dots, m \quad j = 1, \dots, n)$$

In diesem Fall hat man:

$$\Delta f = f'(x) \cdot \Delta x + o(\Delta x), \text{ für } \Delta x \to 0.$$

Oder anders formuliert:

$$f(x+\Delta x) = f(x) + f'(x) \cdot \Delta x + o(\Delta x),$$
$$\text{für } \Delta x \to 0.$$

Oder so formuliert:

$$f(x) = f(x^*) + f'(x^*) \cdot (x - x^*) + o(x - x^*),$$
$$\text{für } x \to x^*.$$

Bei $m=1$ ist $f'(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)$ – eine Zeile

und der Gradient

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad \text{– eine Spalte}$$

(kein großer Unterschied).

Bsp $f(x_1, x_2) = x_1 \cos(x_2 - x_1^2)$

$$\boxed{\begin{matrix} x_1 \rightsquigarrow x \\ x_2 \rightsquigarrow y \end{matrix}} \qquad f(x,y) = x\cos(y - x^2).$$

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

$$\frac{\partial f}{\partial x} \underset{\substack{\uparrow \\ \text{Produktregel}}}{=} \cos(y - x^2) + x \cdot (2x \cdot \sin(y - x^2))$$

$$\frac{\partial f}{\partial y} = -x \sin(y - x^2)$$

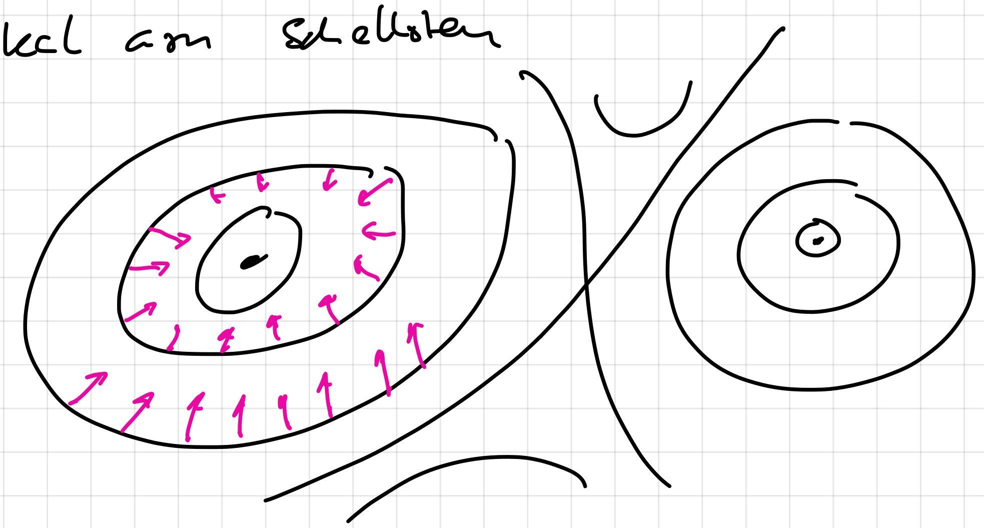
Wir setzen $(x^*, y^*) = (1, -1)$

$$f(x,y) \approx f(x^*, y^*) + \frac{\partial f}{\partial x}(x^*, y^*) \cdot (x - x^*) + \frac{\partial f}{\partial y}(x^*, y^*) \cdot (y - y^*)$$

$$= \left\langle \nabla f(x^*, y^*), \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x^* \\ y^* \end{pmatrix} \right\rangle$$

bei: $x \approx x^*$,
$y \approx y^*$

In Richtung von $\nabla f(x^*, y^*)$ steigt man lokal am schnellsten



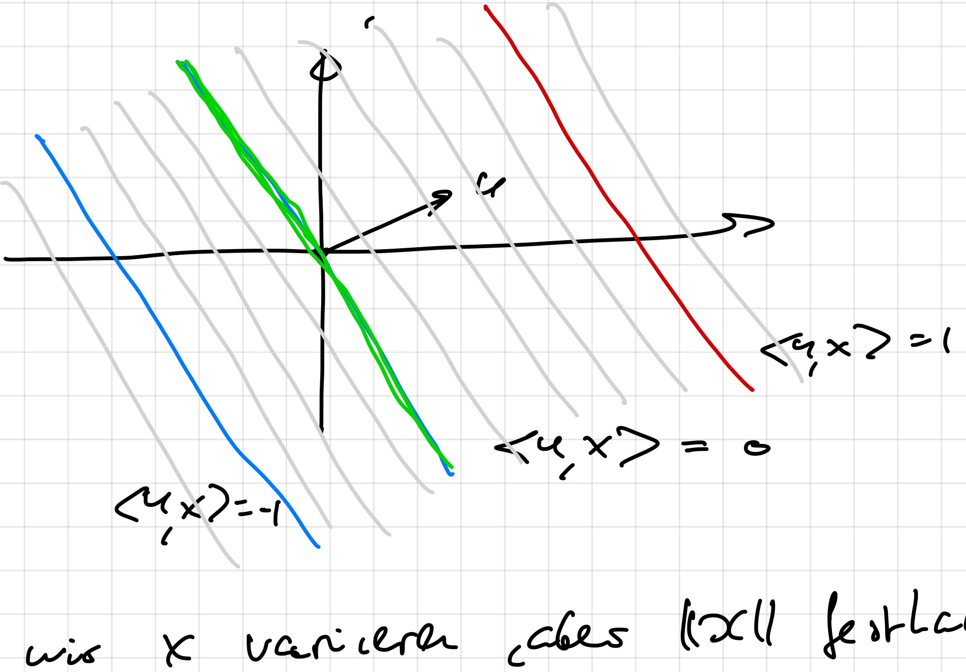
Bemerkung zum Gradienten

Cauchy Schwarz besagt:

$$|\langle u, x\rangle| \leq \|u\| \cdot \|x\|$$

mit der Gleichung genau dann wenn
x und u linear unabhängig sind.
$(x, u \in \mathbb{R}^n)$.

$$-\|u\| \cdot \|x\| \leq \langle u, x\rangle \leq \|u\| \cdot \|x\|$$



Wenn wir x variieren, aber $\|x\|$ festhalten, so kriegen wir bei $\langle u, x\rangle$ den größten Wert, wenn x die Richtung von u hat, z.B. $x = \frac{u}{\|u\|}$.

Und wir kriegen den niedrigsten Wert, wenn x die Richtung von $-u$ hat.

$\rightsquigarrow$ ∇f ist die Richtung des schnellsten Anstiegs von f (lokal)

$-\nabla f$ ist die Richtung des schnellsten Abstiegs von f (lokal).

Außerdem ist $\nabla f(p)$ (lokal) senkrecht zur Niveau-Fläche $\{x : f(x) = f(p)\}$

(p – der fixierte Punkt).

Kettenregel

$f(g(x))$ muss abgeleitet werden.

$$\begin{array}{ccccc} \mathbb{R}^k & \xleftarrow{f} & \mathbb{R}^m & \xleftarrow{g} & \mathbb{R}^n \\ \cup & & \cup & & \cup \\ f(y) = f(g(x)) & & y = g(x) & & x \end{array}$$

$$\underbrace{f(g(x))'}_{\boxed{k \times n}} = \underbrace{f'(g(x))}_{\boxed{k \times m}} \cdot \underbrace{g'(x)}_{\boxed{m \times n}}$$

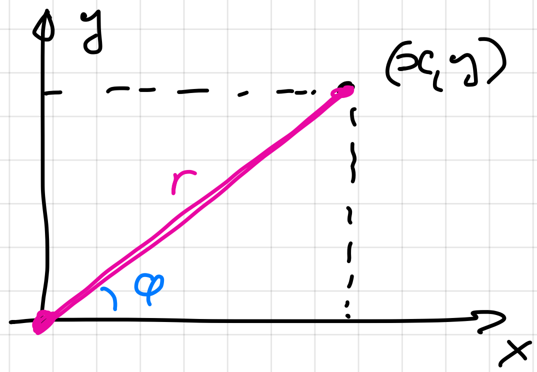
Oder so:

$$J_{f \circ g}(x) = J_f(g(x)) \cdot J_g(x)$$

Bsp.

$$f(x,y) = x\cos(y - x^2).$$

$$g(r,\varphi) = \begin{bmatrix} r\cos\varphi \\ r\sin\varphi \end{bmatrix}$$



$$f = x\cos(y - x^2)$$

$$x = r\cos\varphi$$

$$y = r\sin\varphi$$

$$\left\{\begin{aligned} \frac{\partial f}{\partial r} &= \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial r} \\ \frac{\partial f}{\partial \varphi} &= \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial \varphi} + \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial \varphi} \end{aligned}\right.$$

$$\Updownarrow$$

$$\begin{bmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \varphi} & \frac{\partial y}{\partial \varphi} \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

Hier Bezeichnungen vernachlässigt: x, y sind gleichzeitig Variablen und Funktionen. f hat gleichzeitig zwei Bedeutungen.

Aber man schreibt so in der Physik (z.B.).

$$\frac{\partial x}{\partial r} = \frac{\partial}{\partial r} r\cos\varphi = \cos\varphi$$

$$\frac{\partial y}{\partial r} = \frac{\partial}{\partial r} r\sin\varphi = \sin\varphi$$

$$\frac{\partial x}{\partial \varphi} = \frac{\partial}{\partial \varphi} r\cos\varphi = -r\sin\varphi$$

$$\frac{\partial y}{\partial \varphi} = \frac{\partial}{\partial \varphi} r\sin\varphi = r\cos\varphi$$

$$\Rightarrow \quad \frac{\partial f}{\partial r} = \frac{\partial f}{\partial x} \cdot \cos\varphi + \frac{\partial f}{\partial y} \cdot \sin\varphi$$

$$\frac{\partial f}{\partial \varphi} = \frac{\partial f}{\partial x} \cdot (-r\sin\varphi) + \frac{\partial f}{\partial y} \cdot (r\cos\varphi).$$

$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ haben wir bereits ausgerechnet.

$f(x,y)$

$$x = r\cos\varphi \qquad x(r,\varphi)$$
$$y = r\sin\varphi \qquad y(r,\varphi)$$

$$f(r\cos\varphi, r\sin\varphi) = f(x,y) = f(x(r,\varphi), y(r,\varphi))$$

$$f(x,y) \longleftarrow \begin{pmatrix} x \\ y \end{pmatrix} \longleftarrow \begin{pmatrix} r \\ \varphi \end{pmatrix}$$

$$\frac{\partial}{\partial r} f(x,y) = \left(\frac{\partial f}{\partial x}\right)(x,y) \cdot \frac{\partial x}{\partial r} + \left(\frac{\partial f}{\partial y}\right)(x,y) \cdot \frac{\partial y}{\partial r}$$
$$\frac{\partial}{\partial \varphi} f(x,y) = \left(\frac{\partial f}{\partial x}\right)(x,y) \cdot \frac{\partial x}{\partial \varphi} + \left(\frac{\partial f}{\partial y}\right)(x,y) \cdot \frac{\partial y}{\partial \varphi}$$

Das Gleiche mit Matrizen

$$\underbrace{\begin{bmatrix} \frac{\partial}{\partial r} f(x,y) & \frac{\partial}{\partial \varphi} f(x,y) \end{bmatrix}}_{\dfrac{\partial f(x,y)}{\partial (r,\varphi)}} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x}(x,y) & \frac{\partial f}{\partial y}(x,y) \end{bmatrix}}_{\dfrac{\partial f}{\partial (x,y)}(x,y)} \cdot \underbrace{\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{bmatrix}}_{\dfrac{\partial (x,y)}{\partial (r,\varphi)}}$$

$$g(r,\varphi) := \begin{bmatrix} x(r,\varphi) \\ y(r,\varphi) \end{bmatrix}$$

$$f \circ g = f(x,y).$$

$$\begin{bmatrix} \frac{\partial f(x,y)}{\partial r} & \frac{\partial f(x,y)}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} \cos(y-x^2)+2x^2\sin(y-x^2) & -x\cdot\sin(y-x^2) \end{bmatrix} \cdot \underbrace{\begin{bmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{bmatrix}}_{\substack{\frac{\partial(x,y)}{\partial(r,\varphi)} \\ \| \\ J_g(r,\varphi)}}$$

$x = r\cos\varphi$

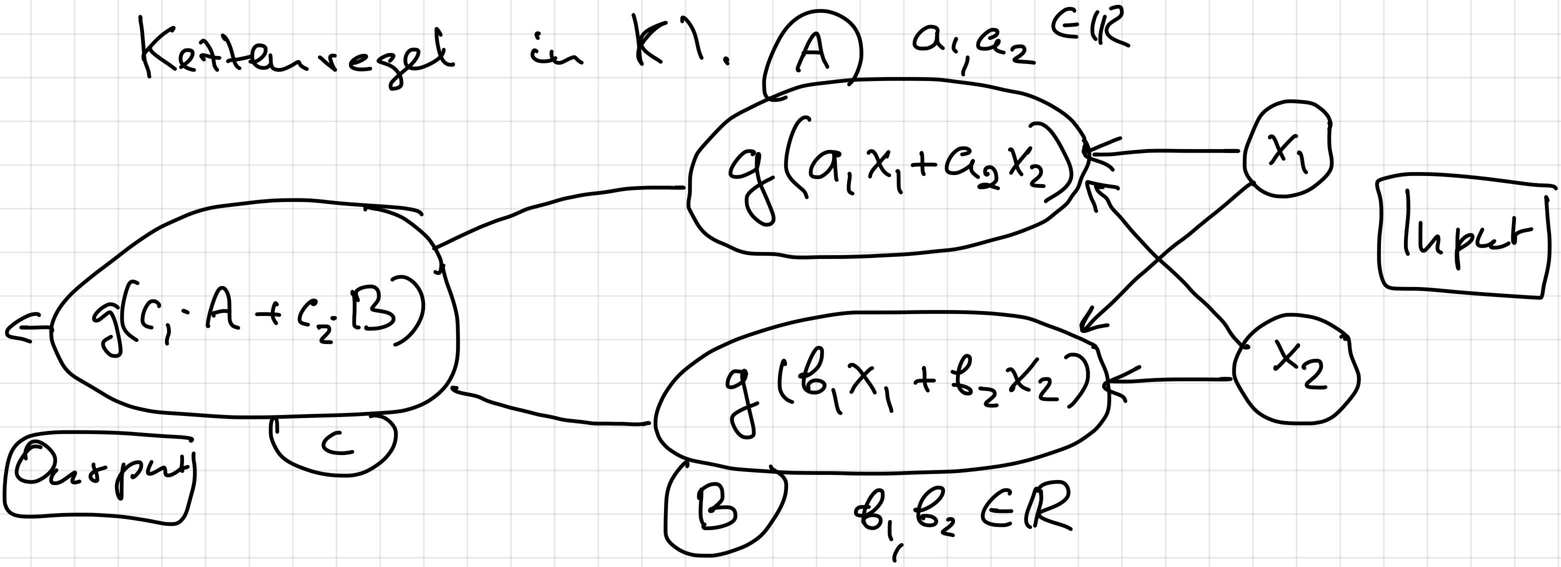
$y = r\sin\varphi$

$$f(x,y) = x\cos(y-x^2)$$

$$\frac{\partial f}{\partial x} = \cos(y-x^2) + 2x^2\sin(y-x^2)$$

$$\frac{\partial f}{\partial y} = -x\sin(y-x^2)$$

Kettenregel in KI.

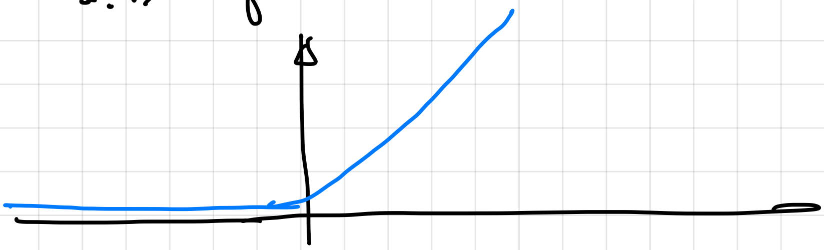


$$A = g(a_1x_1 + a_2x_2)$$

$$B = g(b_1x_1 + b_2x_2)$$

$$C = g(c_1\cdot A + c_2\cdot B)$$

g könnte z.B. $g(t) = \ln(1+e^t)$



$$A = A(a_1, a_2; x_1, x_2)$$
$$B = B(b_1, b_2; x_1, x_2)$$
$$C = C(a_1, a_2, b_1, b_2, c_1, c_2; x_1, x_2).$$

$X^{(i)} = (x_1^{(i)}, x_2^{(i)})$ mit $i = 1, \dots, N$

Stichprobe, mit Labels $L^{(i)} \in \mathbb{R}$

(der gewünschte Output.

Training ist:

$$\underbrace{\min_{a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}} \sum_{i=1}^{N} \Big(\overbrace{C(\underbrace{a_1, a_2, b_1, b_2, c_1, c_2}_{\text{Gewichte}}; x_1^{(i)}, x_2^{(i)})}^{= C^{(i)}} - L^{(i)} \Big)^2}_{\text{Loss function}}$$

$$f(a_1, a_2, b_1, b_2, c_1, c_2)$$

Wir brauchen $\dfrac{\partial f}{\partial (a_1, a_2, b_1, b_2, c_1, c_2)}$ um

einen Gradienten-Abstieg zum Optimieren

zu nutzen.

$$\frac{\partial f}{\partial a_1} = \sum_{i=1}^{N} \frac{\partial}{\partial a_1} \left(C^{(i)} - L^{(i)}\right)^2 = \sum_{i=1}^{N} 2\left(C^{(i)} - L^{(i)}\right) \cdot \frac{\partial C^{(i)}}{\partial a_1}$$

$$\frac{\partial C}{\partial a_1} = \frac{\partial}{\partial a_1} g(c_1 A + c_2 B)$$

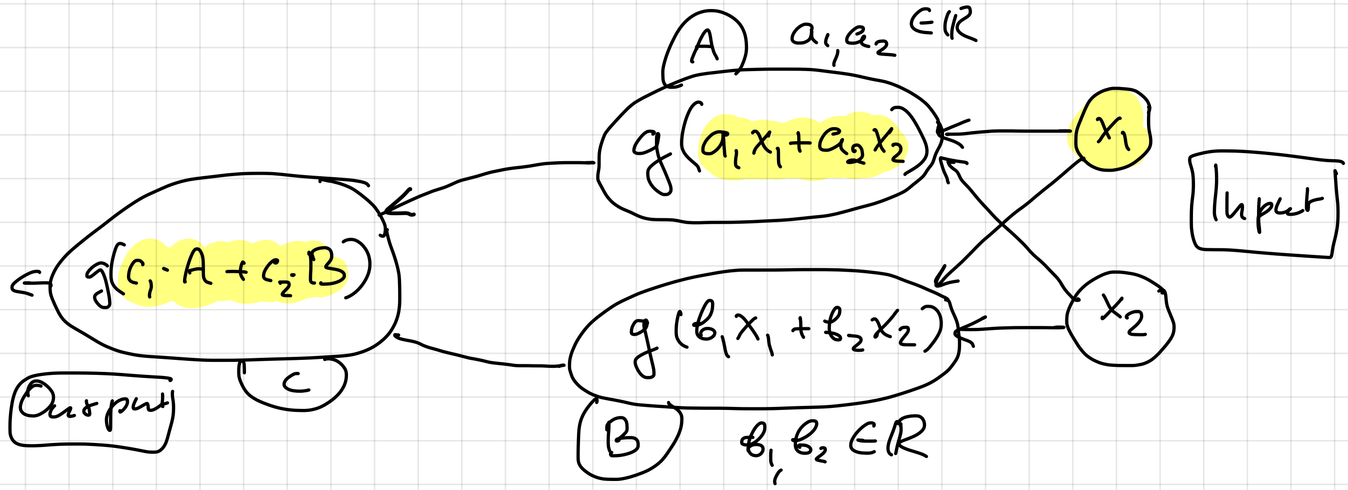
$$= g'(c_1 A + c_2 B) \cdot \frac{\partial (c_1 A + c_2 B)}{\partial a_1}$$

$$= g'(c_1 A + c_2 B) \cdot c_1 \frac{\partial A}{\partial a_1}$$

$$= g'(c_1 A + c_2 B) \cdot c_1 \frac{\partial}{\partial a_1} g(a_1 x_1 + a_2 x_2)$$

$$= g'(c_1 A + c_2 B) \cdot c_1 \, g'(a_1 x_1 + a_2 x_2) \cdot \frac{\partial}{\partial a_1} (a_1 x_1 + a_2 x_2)$$

$$= g'(c_1 A + c_2 B) \cdot c_1 \, g'(a_1 x_1 + a_2 x_2) \cdot x_1$$



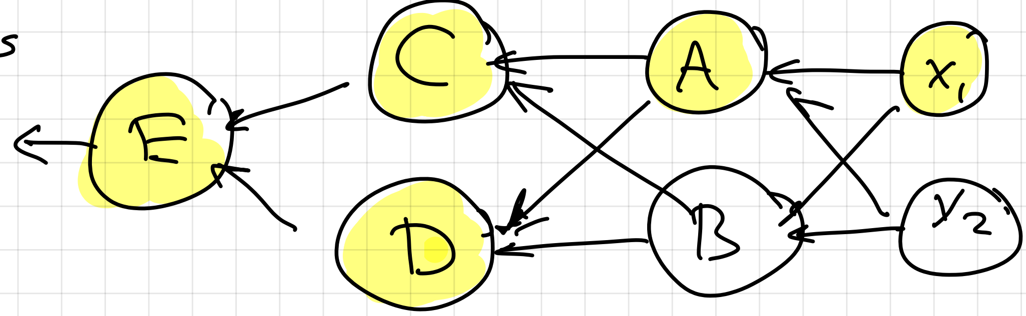
$$\frac{\partial C}{\partial a_2}, \quad \frac{\partial C}{\partial b_1}, \quad \frac{\partial C}{\partial b_2}$$ — Analog.

$$\frac{\partial C}{\partial c_1} = \frac{\partial}{\partial c_1} g(c_1 \cdot A + c_2 \cdot B) =$$

$$= g'(c_1 A + c_2 B) \cdot \frac{\partial}{\partial c_1} (c_1 A + c_2 B)$$

$$= g'(c_1 A + c_2 B) \cdot A$$

Kompliziertes Beispiel

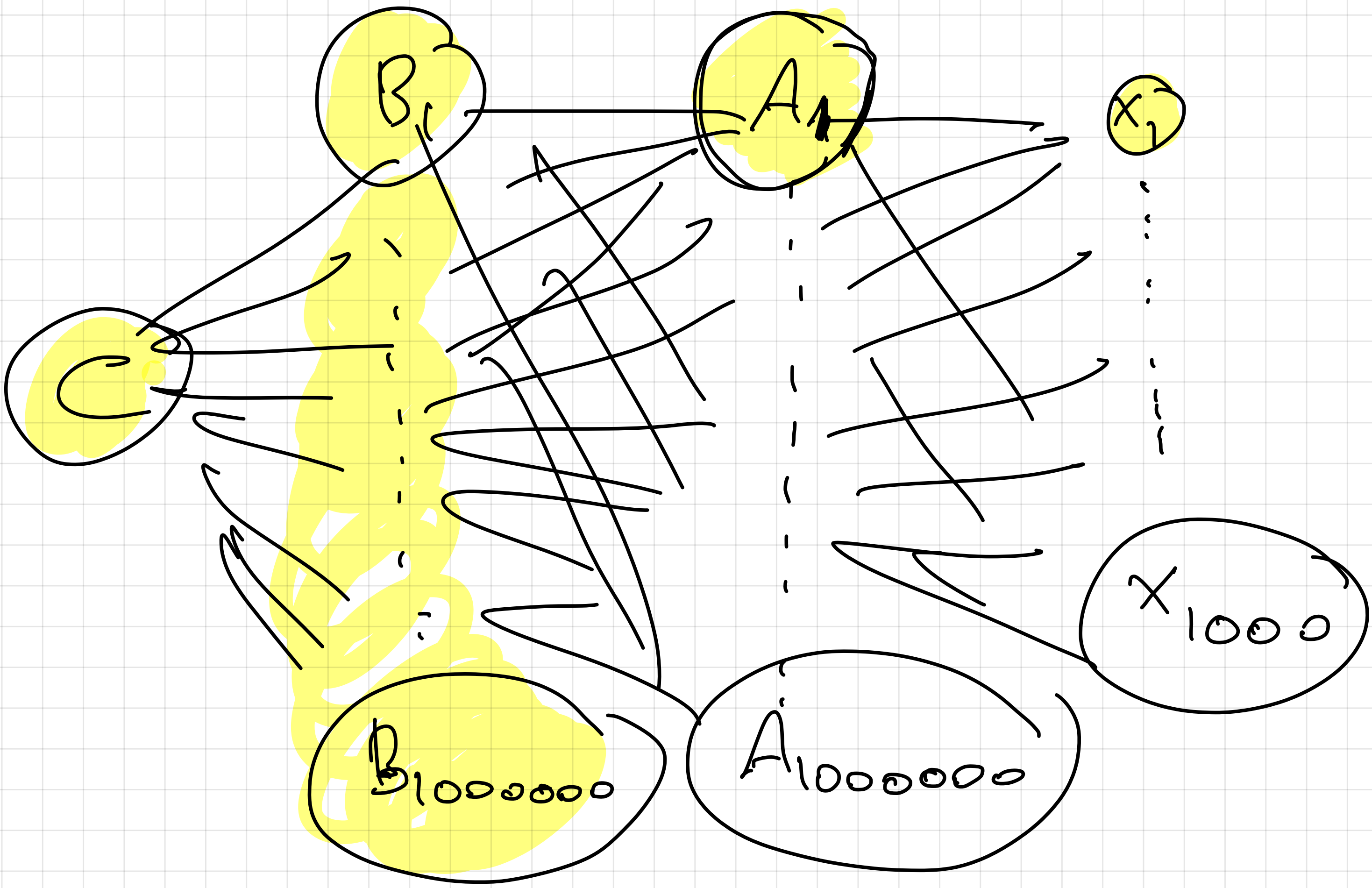


$$\frac{\partial E}{\partial a_1} = \frac{\partial}{\partial a_1} g(e_1 C + e_2 D)$$

$$= g'(e_1 C + e_2 D) \cdot \frac{\partial}{\partial a_1} (e_1 C + e_2 D)$$

$$= g'(e_1 C + e_2 D) \cdot \left(e_1 \frac{\partial C}{\partial a_1} + e_2 \frac{\partial D}{\partial a_1}\right)$$

Noch ein etwas komplizierteres Beispiel



$$A_1 = g(a_{11} x_1 + \dots + a_{1,1000} x_{1000})$$

$$\frac{\partial C}{\partial a_{11}} = g'\left(\sum_{i=1}^{10^6} c_i B_i\right) \cdot \left(\sum_{i=1}^{10^6} c_i \frac{\partial B_i}{\partial a_{11}}\right)$$

$$= g'\left(\sum_{i=1}^{10^6} c_i B_i\right) \cdot \sum_{i=1}^{10^6} c_i \frac{\partial}{\partial a_{11}} g\left(\sum_{j=1}^{10^6} b_{ij} A_j\right)$$

$$= g'\left(\sum_{i=1}^{10^6} c_i B_i\right) \sum_{i=1}^{10^6} c_i g'\left(\sum_{j=1}^{10^6} b_{ij} \cdot A_j\right) b_{i1} \frac{\partial A_1}{\partial a_{11}}$$

$$= g'\left(\sum_{i=1}^{10^6} c_i B_i\right) \sum_{i=1}^{10^6} c_i g'\left(\sum_{j=1}^{10^6} b_{ij} \cdot A_j\right) b_{i1} g'\left(\sum_{k=1}^{10^6} a_{1k} x_k\right) x_1$$