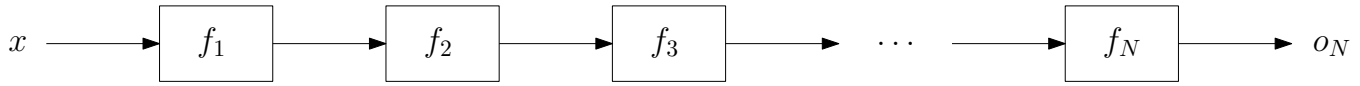
Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **9** — 24 Punkte

Aufgabe 1. *Skip Connections* 7 P.

Gegeben sei ein neuronales Netz mit N linearen Layern, das auf skalaren Eingabedaten $x_i \in \mathbb{R}$ operiert.



Formal bedeutet dies, dass für jeden Layer $i = 1, \ldots, N$,

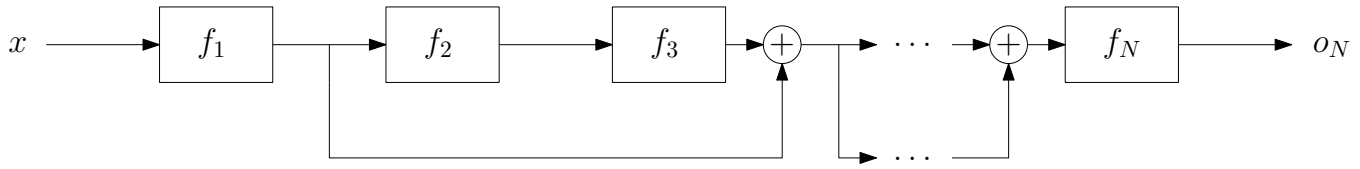
$$s_i = f_i(o_{i-1}) = w_i o_{i-1} + b_i \tag{1}$$
$$o_i = \sigma(s_i) \ , \tag{2}$$

wobei σ eine (beliebige) Aktivierungsfunktion und $o_0 = x$ ist. Einfachheitshalber besteht jeder Layer aus nur einem Neuron, sodass $w_i, b_i, o_i, s_i \in \mathbb{R}$ skalarwertig sind.

(a) Bestimmen Sie die Ableitung $\frac{\partial o_n}{\partial w_1}$ des Outputs nach dem Gewicht des ersten Layers in Abhängigkeit von s_i, w_i (für $i = 1, \ldots, N$), x und der Ableitung der Aktivierungsfunktion $\sigma'(\cdot)$.

(b) Erklären Sie mithilfe von (a) das Vanishing-, bzw. Exploding-Gradient Problem.

Wir ändern nun die Architektur durch das Einführen von Skip Connections, die jeweils eine Kombination an Layern f_{2j}, f_{2j+1} mit geradem und dann ungeradem Index überspringen. Die Anzahl N der linearen Layer sei gerade.



(c) Adaptieren Sie die Formeln (1) und (2) auf die geänderte Architektur.

(d) Bestimmen Sie $\frac{\partial o_n}{\partial w_1}$ für die geänderte Architektur.

(e) Wie wirken sich die Shortcuts bzgl. des Vanishing-, bzw. Exploding-Gradient Problems aus?

Aufgabe 2. *Initialisierungen* 12 P.

Wir betrachten die Initialisierung von linearen Layern in einem neuronalen Netzwerk. Es sei $\mathbf{x} \in \mathbb{R}^{n_{\text{in}}}$ der Input des Layers und $\mathbf{y} \in \mathbb{R}^{n_{\text{out}}}$ der Output, wobei $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$ mit $\mathbf{W} \in \mathbb{R}^{n_{\text{out}} \times n_{\text{in}}}$ und $\mathbf{b} \in \mathbb{R}^{n_{\text{out}}}$.
Wir initialisieren die Einträge w_{ij} der Matrix $\mathbf{W}$ zufällig, wobei jeder Eintrag w_{ij} unabhängig von der gleichen Verteilung gezogen wird. Diese Verteilung der w_{ij} habe Erwartungswert 0, Varianz σ^2 und sei unabhängig von der Verteilung des Inputs $\mathbf{x}$. Der Bias $\mathbf{b}$ sei auf $\mathbf{0}$ initialisiert. Außerdem nehmen wir an, dass Verteilungen der Koordinaten x_j gemeinsam unabhängig sind mit $\mathbb{E}[x_j] = 0$ und $\mathbb{V}[x_j] = \gamma^2$.

(a) Berechnen Sie die Erwartungswert $\mathbb{E}[y_i]$ der Einträge des Outputs zur Initialisierung.

(b) Berechnen Sie die Varianzen $\mathbb{V}[y_i]$ des Outputs in Abhängigkeit der Varianzen des Inputs γ^2.
Hinweis: Für die Varianz einer Summe von Zufallsvariablen Z_i gilt $\mathbb{V}[\sum_i Z_i] = \sum_i \mathbb{V}[Z_i] + \sum_{j \neq k} \operatorname{Cov}(Z_j, Z_k)$.

(c) Motivieren Sie die Wahl einer Initialisierungsverteilung mit Varianz $\sigma^2 = \frac{1}{n_{\text{in}}}$.

Alternativ wählen wir σ^2 unter Berücksichtigung des Gradienten des Trainingsloss L. Dazu nehmen wir an, dass die partiellen Ableitungen $\frac{\partial L}{\partial y_i}$ jeweils einer Verteilung mit Erwartungswert 0 und Varianz γ^2 folgen. Die Verteilung nach der wir die Gewichte w_{ij} initialisieren sei davon unabhängig.

(d) Nach der mehrdimensionalen Kettenregel gilt dass die Jacobi-Matrix $\mathbf{J}_L(x)$ des Trainingsloss die Gleichung

$$\mathbf{J}_L(\mathbf{x}) = \mathbf{J}_L(\mathbf{y})\ \mathbf{J}_y(\mathbf{x})$$

erfüllt.

Drücken Sie diese Gleichung durch die Gradienten $\nabla_x L, \nabla_y L$ und die Matrix $\mathbf{W}$ aus.

Hinweis, die Einträge der Jacobi Matrix einer differenzierbaren Funktion $f : \mathbb{R}^k \to \mathbb{R}^l, \mathbf{z} \mapsto f(\mathbf{z})$ sind definiert als $[\mathbf{J}_f(\mathbf{z})]_{ij} = \frac{\partial f_i}{\partial z_j}$

(e) Berechnen Sie die Erwartungswerte $\mathbb{E}[\frac{\partial L}{\partial x_i}]$ der partiellen Ableitungen.

(f) Berechnen Sie die Varianzen $\mathbb{V}[\frac{\partial L}{\partial x_i}]$ in Abhängigkeit von γ^2.

(g) Motivieren Sie die Wahl einer Initialisierungsverteilung mit Varianz $\sigma^2 = \frac{1}{n_{\text{out}}}$.

Aufgabe 3. *Xavier Initialisierung* 5 P.

In dem Setting von Aufgabe 2 nennt man die Wahl $\sigma^2 = \frac{2}{n_{\text{in}}+n_{\text{out}}}$ Xavier Initialisierung.

(a) Motivieren Sie diese Wahl mithilfe der Ergebnisse aus Aufgabe 2.

(b) Wir initialisieren $\mathbf{W}$ mithilfe einer Normalverteilung $\mathcal{N}(m, s^2)$. Welche Parameter m, s^2 entsprechen einer Xavier Initialisierung.

(c) Wir initialisieren $\mathbf{W}$ mithilfe einer stetigen Gleichverteilung $\mathcal{U}_{[a,b]}$ auf dem Interval $[a, b]$. Welche Intervallgrenzen a, b entsprechen einer Xavier Initialisierung.