# EWI-Hausi-09

## Team AI

## Aufgabe 1:

1a) Sei $X_{ki} \sim \mathcal{N}(0,1)$ jeweils die ZV, welche $w_{ki}$ beschreibt.

$$\Rightarrow \sum_{k=1}^{m} x_k X_{ki} + \underbrace{b_k}_{=0} = \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} X_{ki} \sim \mathcal{N}\Big(\sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} 0, \sqrt{\sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} 1}\Big)$$

$$= \mathcal{N}\Big(0, \sqrt{\lfloor \tfrac{m}{2} \rfloor}\Big) = \mathcal{N}(0, \sqrt{150})$$

b) Durch die breite Varianz entstehen eventuell dendritische Potential weit weg von 0. ~~Solche~~ Dadurch würde das Problem von vanishing gradient auftauchen.

c) $X_{ki} \sim \mathcal{N}(0, n_{in}^{-\frac{1}{2}})$

$$\sum_{k=1}^{m} x_k X_{ki} + b_k = \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} X_{ki} \sim \mathcal{N}\Big(0, \Big(\sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} \Big(\frac{1}{\sqrt{n_{in}}}\Big)^2\Big)^{\frac{1}{2}}\Big) = \mathcal{N}\Big(0, \Big(\frac{m}{2 n_{in}}\Big)^{\frac{1}{2}}\Big)$$

$$= \mathcal{N}\Big(0, \Big(\frac{1}{2}\Big)^{\frac{1}{2}}\Big)$$

d) Die W'keit, dass die dendritischen Potentiale jetzt hohe Werte annehmen, ist deutlich geringer.

2.) I entspricht A und II entspricht B.
Denn der tanh staucht die breite Verteilung, die durch I verursacht wird, an den Rändern. Wodurch bei A viele $\sim 1$ und $\sim -1$ entstehen. Außerdem ist die Ableitungen an diesen Stellen fast 0, was zur Abbildung der Gradienten passt. Die Abbildung passt auch zu denn berechneten Verteilungen

3) Die ~~Xavier-Initialisierung~~ skalierte Normalverteilung reicht das Problem der Standardnormalverteilung einfach nur eine Schicht weiter.
Die Xavier-Initialisierung löst dieses Problem und pro Schicht wird dadurch unter Beachtung aller Neuronen, die Varianz beibehalten.

---

## Aufgabe 2:

1a) Zunächst: $\frac{\partial}{\partial w} \frac{\lambda}{2} \sum_{\ell=1}^{L} \| W^{\ell} \|_F^2 = \frac{\lambda}{2} 2w = \lambda \cdot w$

$$\Rightarrow w(t+1) = w(t) - \eta \nabla E(w(t))$$
$$= w(t) - \eta \nabla E_0(w(t)) - \eta \frac{\partial}{\partial w} \frac{\lambda}{2} \sum_{\ell=1}^{L} \| W^{\ell} \|_F^2$$
$$= w(t) - \eta \nabla E_0(w(t)) - \eta \lambda w(t)$$
$$= (1 - \lambda\eta) w(t) - \eta \nabla E_0(w(t))$$

---

b) i) Mit den Werten eingesetzt ergibt sich:

$$w(t+1) = w(t)(1 - 0{,}8 \cdot 0{,}5) - 0$$
$$= 0{,}6 \cdot w(t)$$

$$\Rightarrow w(t) = (0{,}6)^t \cdot 2 \qquad \Rightarrow w(10) \approx 0{,}012$$

ii) Das Gewicht sinkt logarithmisch.

iii) $\lim_{t \to \infty} w(t) = \lim_{t \to \infty} (0{,}6)^t \cdot 2 = 0$ (da streng mon. fallend und nach unten beschränkt)

iv) In diesem Fall wird jede Iteration durch die Ableitung beeinflusst. Also wird das Gewicht immer verändert, solang es nicht optimal ist.

---

c) Regularisierung hilft vanishing gradients einzudämmen, da jedes Gewicht natürlich ~~veränd~~ verkleinert wird (absolut) und nach einigen Iterationen sollten die meisten Gewichte nahe um 0 gestreut sein.

2.) Wie man auf Blatt 3 sieht, fällt der Faktor des Inputs in der Ableitung nach b. weg. D.h. der Input hat deutlich geringeren Einfluss auf die Anpassung des Bias und daher auch verzerrte Inputs.

3.)