Blatt 9

Luca Krüger, Jonas Otto, Jonas Merkle (Gruppe R)

8. Juli 2019

1 Gewichtsinitialisierung

1. a) $u_i^{(1)}$ ist normalverteilt mit $\mathcal{N}(0, \sqrt{\frac{m}{2}})$.

   b) Durch die Verwendung von tanh als Transferfunktion ändert sich die Ausgabe der Neuronen bei Änderung des bereits relativ großen u_i kaum, die Ableitung $f'(u_i) \to 0$. Die Errorfunktion ist wie in a) ebenfalls über die Summe über die Gewichte $w_1, \ldots, w_n$ definiert. Dies führt dann zu großen Lernschritten, da gilt $E(w) \propto \mathcal{N}(0, \sqrt{\frac{m}{2}})$

   c) $u_i^{(1)}$ ist normalverteilt mit $\mathcal{N}(0, \frac{1}{\sqrt{2}})$.

   d) Alle Potentiale u_i liegen in einem Bereich, in dem $f'(u)$ deutlich von 0 verschieden ist. Die Gewichte werden dadurch nicht mit Extremwerten initialisiert, die statistisch zwar selten, aber eben doch auftreten können.

2. Verteilung A passt zu Strategie I, da in Verteilung A die Gradienten alle ≈ 0 sind und die axonalen Potentiale 1 oder -1, was auf ein betragsmäßig großes Argument der tanh Aktivierungsfunktion hindeutet. Verteilung B entspricht der skalierten Normalverteilung. Die Gradienten sind von 0 verschieden und die axonalen Potentiale nahezu gleichverteilt.

2 Regularisierung

1. Einfluss auf Lernregeln

   a)

$$\begin{aligned} \frac{dE}{dw} &= \nabla E_0\left(w(t)\right) + \lambda w(t) \\ \implies w(t+1) &= w(t) - \eta \nabla E_0\left(w(t)\right) - \eta\lambda w(t) \\ &= (1 - \eta \cdot \lambda) w(t) - \eta \cdot \nabla E_0\left(w(t)\right) \end{aligned}$$

   b) i. $w(n) = 2 \cdot 0.6^n$, $w(10) = 0.012$

ii. Das Gewicht nimmt exponentiell ab.

iii. Für $\eta \cdot \lambda < 1$ gilt:

$$\lim_{t \to \infty} w(t) = \lim_{t \to \infty} w(0) \cdot (\eta \cdot \lambda)^t = 0$$

iv. $\nabla E_0(w(t)) \neq 0$ gilt auch für $t \to \infty$, somit gilt keineswegs $\lim_{t \to \infty} w(t) = 0$

c) Durch den Regularisierungsterm werden die Gewichte in jedem Schritt reduziert, auch wenn $\nabla E \approx 0$ gilt. Dadurch bewegen sich die suboptimal initialisierten Gewichte schnell in sinnvollere Bereiche.

2. Der Term für das Update des Bias enthält im Gegensatz zum Term für das Gewicht den Faktor x_μ nicht. Verrauschter Input wirkt sich also bereits mit der bekannten Lernregel nicht direkt auf den Bias aus. Starkes Rauschen wird sogar durch $f(wx + b) \cdot f'(wx + b)$ geglättet.

3. Mit größerem λ nähert sich die Errorfunktion $|w|^2$ an:

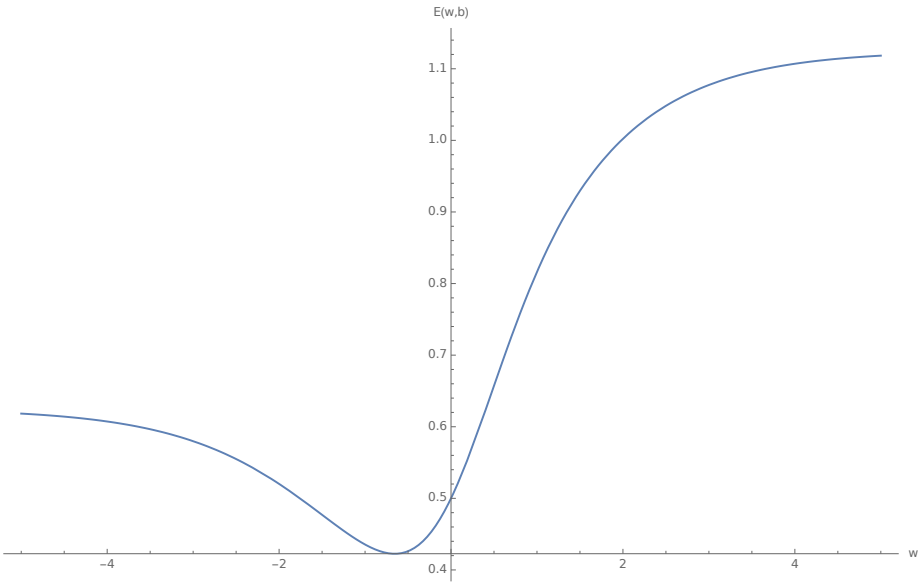

Abbildung 1: Errorfunktion mit $\lambda = 0$

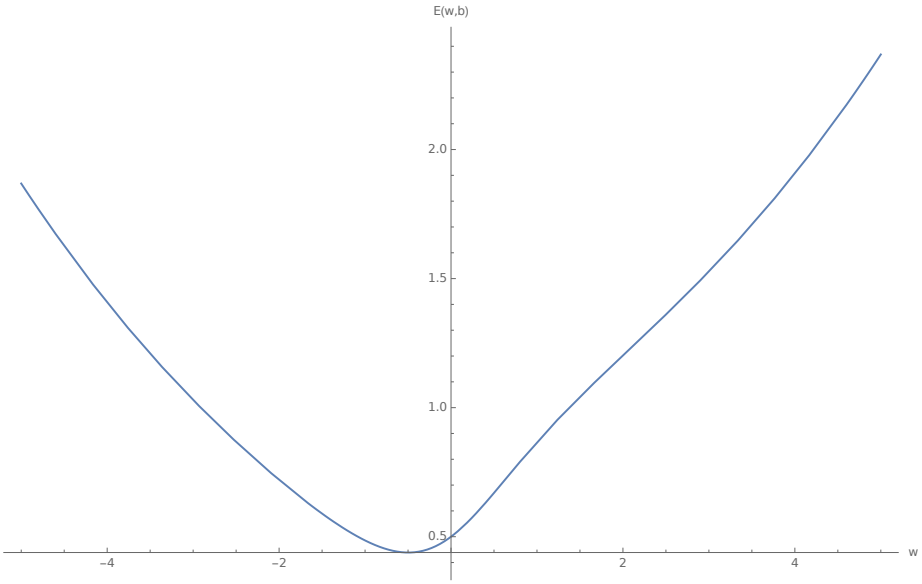

Abbildung 2: Errorfunktion mit $\lambda = 0.2$

4. (Siehe Jupyter Notebook)