Aufgabe 2:

2a)

	$\sqrt{s(t)+\varepsilon} < 1$	$\sqrt{s(t)+\varepsilon} > 1$
vanishing ∇	gemildert	verstärkt
exploding ∇	verstärkt	gemildert

b) $s(t)$ hängt auch direkt von $\nabla E(w(t-1))$ ab.
Das heißt, falls $\nabla E(w(t-1))$ klein ist, dann wahrscheinlich auch $s(t)$. (vanishing gradient)
Und falls $\nabla E(w(t-1))$ groß ist, dann wahrscheinlich auch $s(t)$.
(Und falls $s(t)$ groß/klein dann auch $\sqrt{s(t)+\varepsilon}^{-1}$)

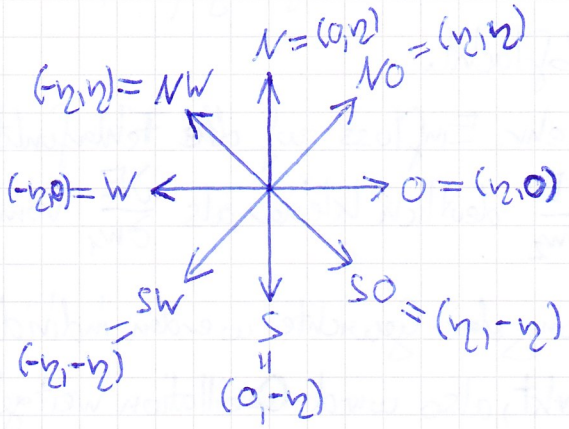
~~Aufgabe 3:~~

3a) $s(t) = \nabla E(w(t-1))^2$

$w(t) = w(t-1) - \eta \cdot \operatorname{sgn}(\nabla E(w-1))$

b) In jeder Iteration wird das Gewicht in Richtung der Ableitung von E angepasst. Und zwar immer mit Schrittweite η.

c)



oder Update ist (0,0) und wir bleiben stehen.

d) Das Verfahren ähnelt jetzt SuperSAB oder RPROP.

4a) $s(2) = \nabla E(w+1))^2 = g_2^2$

$C_a(-2,2) = \left| \frac{g_2}{\sqrt{s(2)+0}} \right| - |g_2| = \left| \frac{g_2}{|g_2|} \right| - |g_2| = 1 - |g_2|$

b) i) Für größer werdendes β wird die Fläche in der Grafik immer "rötlicher". Das heißt an den roten Stellen wird der Gradient g_2 verstärkt, also ist dort der Einfluss von g_1 größer.

ii) Falls beide Gradienten klein sind, ~~sind~~ befinden wir uns mittig in der Grafik und g_2 wird also verstärkt. Dadurch wird das vanishing gradient problem gelindert.
Und am Rand, wo g_1 und g_2 groß sind, wird g_2 abgeschwächt. Also wird hier das exploding gradient problem abgelindert.

5a) Für große β wird die Magnitude stärker beachtet. Das heißt dass der große Fehler am Anfang ein großes s_1 „(langsam) aufbaut" und dadurch, dass β so groß ist wird das s_1 auch langsam wieder „abgebaut". Man kann auch sagen, dass s_1 weniger reaktiv wird für große β und dadurch die Kurve abflacht.

b) Da w_2 mehr Einfluss auf die Fehlerfunktion hat.

Also ist $\frac{\partial E}{\partial w_2}$ deutlich kleiner als $\frac{\partial E}{\partial w_1}$ vom gegebenen Startpunkt.

c) Die Anpassung der Gewichte werden individuell geschwächt und verstärkt, also wird Oszillation weitgehend vermieden. Im Gegensatz zur Momentum-Optimierung, wo ein fixer Momentumfaktor α besteht.

d) Für w_2 sehr klein und w_1 sehr groß gewählt als Startpunkt.