Blatt 8

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

1. Juli 2019

1 Momentum Optimierer

1.

$$\begin{aligned} m(0) &= 0 & w(0) &= 20 \\ m(1) &= 80 & w(1) &= 12 \\ m(2) &= 120 & w(2) &= 0 \end{aligned}$$

2. Beschleunigung/Stabilisierung

	$\frac{\partial E}{\partial w(1)} < 0$	$\frac{\partial E}{\partial w(1)} > 0$
$\frac{\partial E}{\partial w(0)} < 0$	$+$	$-$
$\frac{\partial E}{\partial w(0)} > 0$	$-$	$+$

3. a) $m(t)$ ist durch seine rekursive Definition über $m(t-1)$ und den Gradienten $\frac{\partial E}{\partial w(t-1)}$ eine Differentialgleichung zweiter Ordnung. $m(t)$ ist daher eine (in unserem Fall gedämpfte) Schwingung. Die Gewichte $w(t)$ hängen direkt von $m(t)$ ab. Dadurch ist ein Überschwingen des Pfades von $w(t)$ in Abhängigkeit der gewählten Parameter η, α und der Startwerte $w(0), m(0)$ möglich.

   b) $m(t)$ zeigt wie der Gradient $\frac{\partial E}{\partial w(t-1)}$ in die Richtung des größten Anstieges.
   Die Gewichte $w(t)$ sollen aber in Richtung des kleinsten Anstieges optimiert werden, weshalb $m(t)$ mit negativem Vorzeichen in der Definition von $w(t)$ vorkommt.

   c) Oszillationen treten z.B. auf, wenn der Parameter α groß gewählt wird:

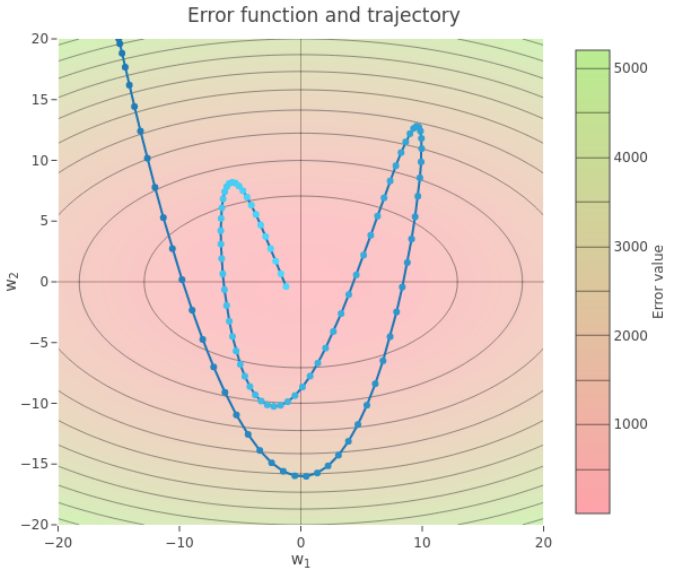

Abbildung 1: Oszillation durch zu großen Momentumfaktor

d) Geeignete Parameter sind z.B. $\eta = 0.02$ und $\alpha = 0.5$, nach 20 iterationen beträgt der Fehler 0.0028:

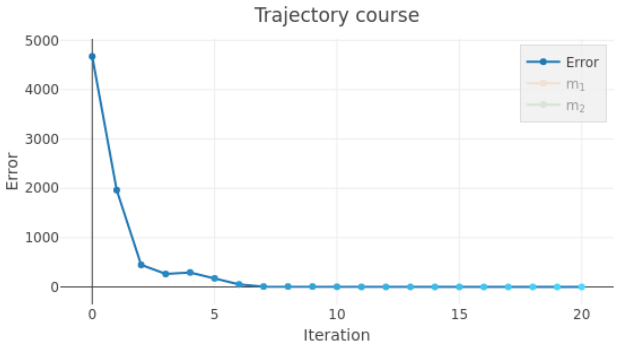

Abbildung 2: Fehler bei geeigneter Parameterwahl

2 Adaptive Learning Rates

1. (siehe Jupyter Notebook)
2. a) Einfluss von $\sqrt{s(t) + \epsilon}$:

Problem	$\sqrt{s(t)+\epsilon} < 1$	$\sqrt{s(t)+\epsilon} > 1$
vanishing gradients	wird gemildert	wird verstärkt
exploding gradients	wird verstärkt	wird gemildert

b) Angenommen, das *exploding gradient* Problem tritt auf, und es gilt $\sqrt{s(t)+\epsilon} < 1$, wird $s(t)$ in nachfolgenden Schritten auch erhöht (da $s(t+1)$ abhängig von $\nabla E(w(t))$), sodass in nachfolgenden Iterationen gilt $\sqrt{s(t)+\epsilon} > 1$, was das *exploding gradient* Problem mildert. Analog im Fall *vanishing gradient*.

3. a) $\beta = 0, \quad \epsilon = 0$

$$
\begin{aligned}
& s(t) = \nabla E(w(t-1))^2 \\
\implies & w(t) = w(t-1) - \eta \frac{\nabla E(w(t-1))}{\sqrt{\nabla E(w(t-1))^2}} \\
\implies & w(t) = w(t-1) - \eta \operatorname{sgn}(\nabla E(w(t-1)))
\end{aligned}
$$

b) (Plot siehe Jupyter Notebook)
In diesem Fall wird w immer nur um η in Richtung des Minimums verändert.

c) Die Updates können sich in jeder Komponente nur um 1, -1 oder 0 ändern. Dies entspricht einem Schritt parallel zu X bzw. Y Achse oder diagonale Bewegung, wobei die Schrittweite in jede Achsenrichtung 1 oder 0 beträgt.

4. a) $\beta = 0$ und $\epsilon = 0$

$$
\begin{aligned}
& \implies s(1) = g_1^2 \quad \wedge \quad s(2) = g_2^2 \\
& \implies C_a(-2, 2) = -1
\end{aligned}
$$

b) i) Die Bedeutung der Magnitude kann anhand der Ausdehnung der Grafik in g_2 Richtung erkannt werden: Für $\beta = 0$ ist neben der Unabhängigkeit von g_1 erkennbar, dass der Betrag des Gradienten normiert wird, die Veränderung beträgt $-(g_2 - 1)$. Für größere β ist eine stärkere Verstärkung kleiner Gradienten erkennbar.

ii) Die Normalisierungseigenschaft lässt vermuten, dass nahe des Ursprungs der Grafik eine Verstärkung des Gradienten und am Rand der Grafik eine Abschwächung zu erkennen ist. Dies bestätigt sich in der Grafik.

5. a) Mit steigendem β, steigt die Gewichtung der Magnitude. Vorherige $s(t-1),\ s(t-2),\ \ldots,\ s(t-j)$ werden also stärker gewichtet, $s(t)$ ändert sich dadurch langsamer.

b) Der Gradient zeigt initial in negative w_2 Richtung, wodurch s_2 ansteigt. Über die weiteren Iterationen bleibt der Gradient in der 2. Komponente betragsmäßig größer, s_2 bleibt also auch größer als s_1 um den Gradienten zu normalisieren.

c) Der Pfad läuft zielstrebiger ins Minimum, da nicht durch große Schrittweiten vom optimalen Pfad abgewichen wird, was beim momentum Verfahren besonders bei großem Parameter α vorkommt.

d) Für den Startpunkt $w = \begin{pmatrix} -17 \\ 1.5 \end{pmatrix}$ liegt s_1 größtenteils über s_2.