Einführung in Neuronale Netze
Backpropagation Learning

Probleme bei Backpropagation

Das Backpropagation-Verfahren basiert bekanntlich auf einem Gradientenabstieg. Obwohl diese Idee naheliegend und relativ einfach ist, beinhaltet dieses Verfahren doch eine Reihe von Problemen.

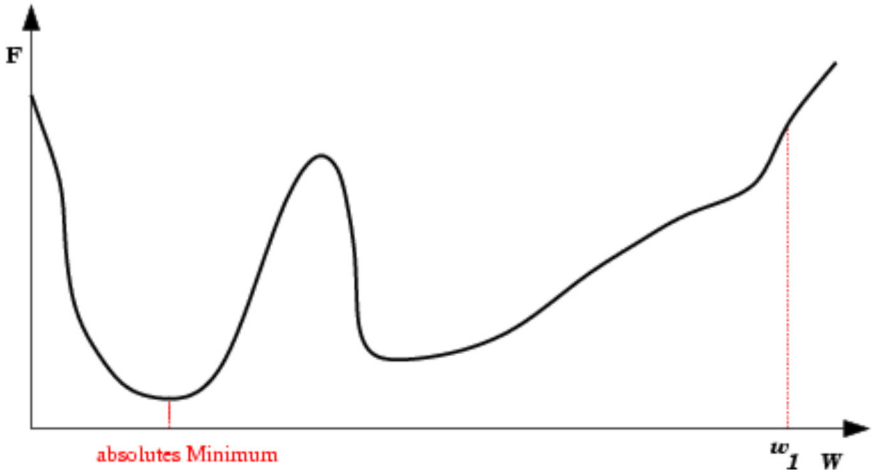
Diese beruhen im wesentlichen darauf, daß dem Verfahren lediglich die lokale Umgebung, eben der Gradient, bekannt ist.

Mit den wesentlichsten Problemen werden wir uns im folgenden Kapitel auseinandersetzen.

Lokale Minima

Schon bei der Einführung der Lernregel konnte nie eindeutig festgelegt werden, ob das Verfahren des Gradientenabstiegs in einem lokalen oder absoluten Minimum endet. Der Grund liegt in der zufälligen Wahl von $\vec{w}$, mit dem man den Gradientenabstieg startet.

Lag $\vec{w}$ in der Nähe eines absoluten Minimums, endet das Gradientenabstiegs-Verfahren in dem optimalen Minimum. Lag $\vec{w}$ hingegen in der Nähe eines lokalen Maximums, führt für dieses Verfahren kein Weg mehr aus dem suboptimalen Minimum heraus.



Durch Betätigen von vorheriger Schritt und nächster Schritt kann schrittweise nachvollzogen werden, warum das Gradientenverfahren zwangsläufig in ein suboptimales Minimum führt, wenn es in w_1 startet.

Die Dimension des Netzes wird durch die Anzahl der Verbindungen zwischen den Neuronen beschrieben. Mit wachsender Dimension des Netzes wird die Fehleroberfläche immer stärker zerklüftet. Mit der Zerklüftung steigt auch die Anzahl der lokalen Minima in der Fehleroberfläche. Damit wird es dem Verfahren weiter erschwert, ein globales Minimum zu erreichen.

Zudem können tiefe Täler in der Fehleroberfläche liegen, die eine relativ geringe Ausdehnung besitzen. Es besteht somit die Möglichkeit, daß dieses Tal bei einer zu großen Lernrate übersprungen wird, obwohl eventuell ein absolutes Minimum in dessen Talsohle liegt.

Symmetry Breaking

Dieses Problem tritt bei vollständig verbunden Feedforward-Netzen auf. Bei dieser Art von Netzen dürfen die Gewichte nicht alle gleich intialisiert werden.

Dies hätte zur Folge, daß alle Gewichte, die von einem Eingabeneuron ausgehen bzw. ein Ausgabeneuron erreichen, auch nach der Gewichtsmodifikation stets den gleichen Wert aufweisen.

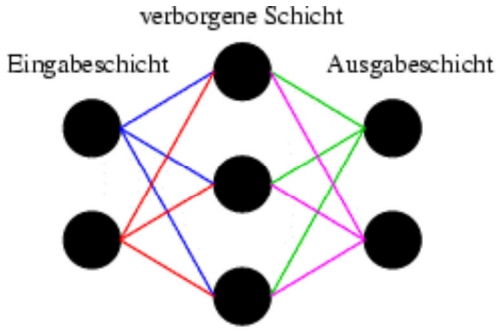
Das Problem kann durch zufällige Initialisierung der Gewichte behoben werden. Verdeutlichen Sie sich das Problem des Symmetry Braking am folgenden Beispiel :

Gegeben sei ein Netz mit zwei Eingabe-Neuronen, drei verborgenen Neuronen und zwei Ausgabe-Neuronen mit exakt gleicher Gewichtsbelegung. Nach dem Forward-Pass für

Die Formel zur Veränderung der Gewichte lautet :

$$\triangle w_{h,i,j} = \eta \, \delta_{h,i} \, o_{h-1,j}.$$

ein Muster *p* liegt somit bei allen verdeckten Neuronen dieselbe Ausgabe vor.



Vor der ersten Gewichtsmodifikation sind alle Gewichte im Netz gleich.

Da die Ausgaben der ***j*** verborgenen Neuronen gleich sind, werden alle Gewichte, die zu dem ***i-ten*** Ausgabeneuron führen, um den selben Faktor verändert. Aber auch die rekursiv berechneten δ der verborgenen Neuronen unterscheiden sich nicht voneinander. Damit werden auch die Gewichte der Verbindungen, die von dem selben Eingabeneuron wegführen, um den gleichen Faktor verändert. Dieser Effekt wird durch die nebenstehende Abbildung erläutert.

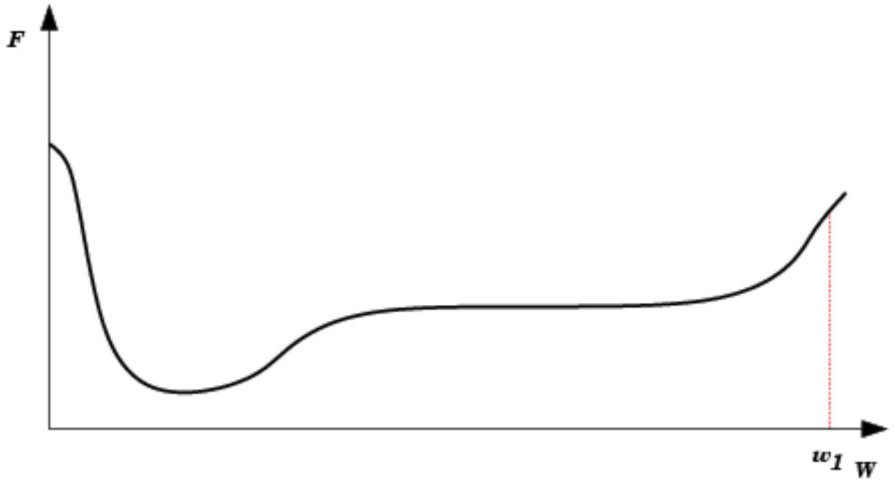
Nach dem Training sind also alle Gewichte der Verbindungen, die von dem selben Eingabeneuron wegführen bzw. zu dem selben Ausgabeneuron führen, gleich. Somit ist auch für jedes weitere Trainingsmuster die Ausgabe aller verborgenen Neuronen gleich und die Symmetrie der Gewichte bleibt erhalten.

Um sicherzustellen, daß Sie den Backpropagation Lernregel wirklich verstanden haben, sollten sie dieses Beispiel nachvollziehen können. Machen Sie sich klar, warum dieses Problem nur bei vollständig verbunden Feedforward-Netzen auftreten kann! Verifizieren sie die Behauptungen mit Hilfe des Applets am Ende des Kapitels.

Flache Plateaus

Die Größe der Gewichtsveränderung wird bei Gradientenabstiegsverfahren maßgeblich durch den Betrag des Gradienten bestimmt.

In flachen Plateaus ist der Gradient sehr klein und das Verfahren stagniert nahezu.



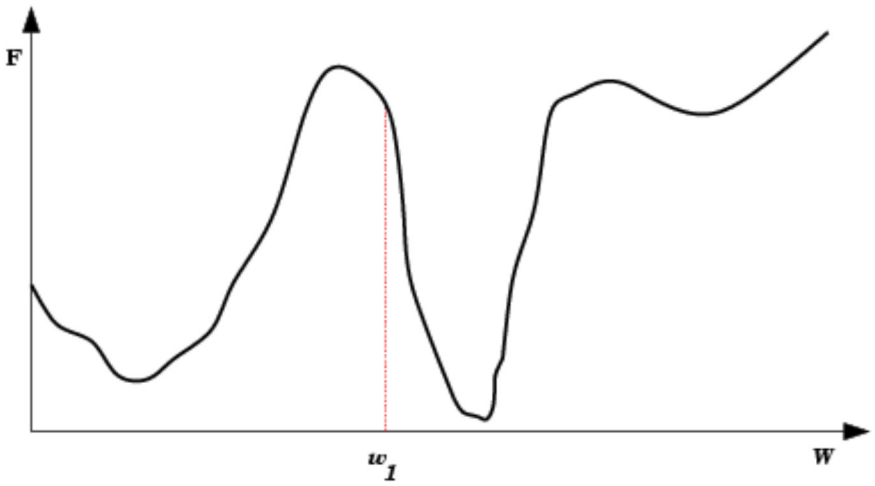
Durch Betätigen von vorheriger Schritt und nächster Schritt wird schrittweise verdeutlicht, daß besonders viele Iterationsschritte benötigt werden, um ein flaches Plateau zu verlassen. Der Gradient kann im Extremfall sogar zum Nullvektor werden, obwohl die sigmoiden Ausgabefunktion diesem Effekt entgegenwirken sollte. In diesem Fall kann das Verhalten des Verfahrens mit dem Erreichen eines Minimums verwechselt werden.

Oszillation

Bei einer unglücklichen Konstellation von Lernrate und Fehleroberfläche kann das Gradientenabstiegsverfahren oszillieren.

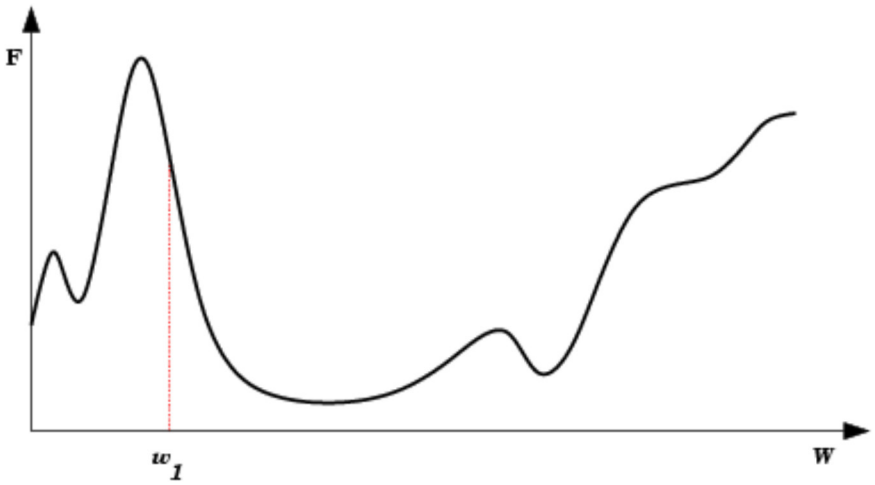
Dieser Fall tritt ein, wenn der Gradient am Rande eines Tals so groß ist, daß die Gewichtsveränderung einen Sprung an die gegenüberliegende Seite des Tals bewirkt.

Ist das Tal dort genauso steil, besitzt der Gradient den gleichen Betrag, aber das umgekehrte Vorzeichen. Also springt das Verfahren zurück zum Ausgangspunkt und das Verfahren oszilliert.



Durch Betätigen von vorheriger Schritt und nächster Schritt wird das direkte Oszillieren schrittweise verdeutlicht. Dieser Effekt tritt vor allem bei relativ steilen Tälern in der Fehleroberfläche auf.

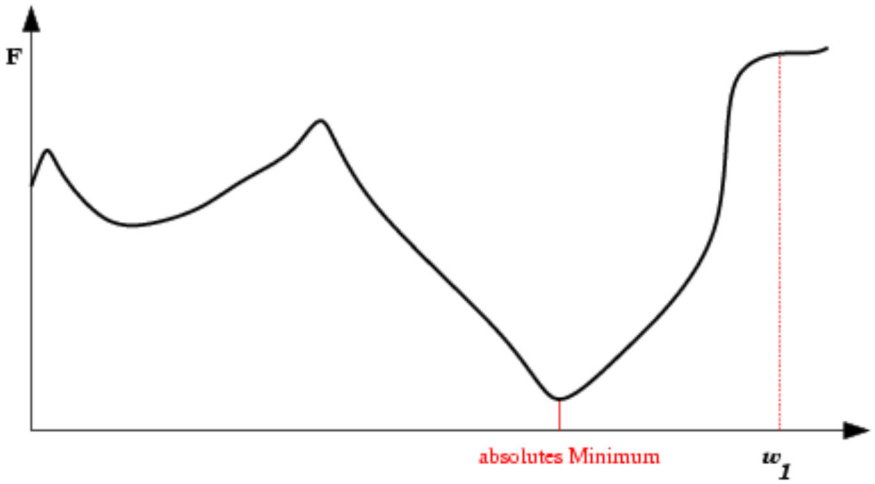
Aber neben dem Problem des direkten Oszillierens kann auch das indirekte Oszillieren auftreten.



Durch Betätigen von vorheriger Schritt und nächster Schritt wird dieser Effekt schrittweise verdeutlicht.

Verlassen guter Minima

Liegt das globale Minimum in einem steilen Tal, kann der Betrag des Gradienten sehr groß werden. In diesem Fall kann die Gewichtsveränderung das Verfahren aus der Umgebung des optimalen Minimums in die Umgebung eines suboptimalen Minimums führen.



Durch Betätigen von vorheriger Schritt und nächster Schritt wird dieser Effekt schrittweise verdeutlicht.

Mögliche Problemlösungen

Den oben aufgezeigten Problemen kann durch verschiedene Maßnahmen begegnet werden.

Die Wahl der Lernrate

Die Wahl der Lernrate η ist entscheidend für das Verhalten des Gradientenabstiegs. Wie wir gesehen haben, kann eine ungeschickte Wahl der Lernrate zu den verschiedensten Problemen führen.

Die optimale Lernrate hängt von vielen Faktoren ab. Zu diesen gehören u.a. das Problem selbst, die Wahl der Trainingsdaten und die Größe und Topologie des Netzes.

Eine geschickte Wahl der Lernrate ist eine gute Möglichkeit, das Problem der lokalen Minima zu bewältigen. Generell bewirkt eine große Lernrate starke Sprünge auf der Fehleroberfläche. Damit wächst das Risiko, ein Minimum in einem engen Tal zu überspringen, welches vielleicht sogar optimal wäre.
Die Wahl einer kleinen Lernrate würde dies verhindern. So wäre sichergestellt, auch auf einer stark zerklüfteten Fehleroberfläche ein für das Problem akzeptables Minimum zu erreichen.
Aber auch dem Problem der Oszillation würde somit begegnet.

Um die Oszillation vollends auszuschließen, müßte die Lernrate allerdings sehr klein gewählt werden, da der Gradient beliebig groß werden kann. Eine zu kleine Wahl der Lernrate hätte aber eine rapide Erhöhung der Trainingszeit zur Folge.
Bei Erreichen eines flachen Plateaus kann die Trainingszeit sogar inakzeptabel groß werden. Flache Plateaus erfordern also eine Erhöhung der Lernrate.

Aber auch die Struktur der Trainingsdaten spielt eine Rolle bei der Wahl der Lernrate. Komplexe Daten werden besser erkannt und generalisiert, wenn die Lernrate klein gehalten wird. Falls die Stützstellen innerhalb des Datensatzes nahe aneinander liegen, hat sich ebenfalls eine kleine Lernrate von Vorteil erwiesen.

Zusammenfassend läßt sich also Folgendes über die Modifikation der Lernrate festhalten:

Vergrößerung der Lernrate und die Folgen

- Ein großes η bewirkt große Sprünge auf der Fehleroberfläche.
- Flache Plateaus werden schneller durchlaufen.
- Minima in steilen Tälern werden übersprungen.
- Die Gefahr der Oszillation nimmt zu.
- Gute Minima können verlassen werden.

Verkleinerung der Lernrate und die Folgen

- Ein kleines η bewirkt kleine Bewegungen auf der Fehleroberfläche.
- Täler werden nicht mehr so leicht übersprungen.
- Komplexe Daten werden besser gelernt.
- Eine große Datendichte wird besser bewältigt.
- Die Umgebung eines Minimums wird nicht mehr verlassen.
- Oszillation wird immer unwahrscheinlicher.
- Der Gradientenabstieg wird zusehens langsamer.
- Auf flachen Plateaus kann das Verfahren nahezu stagnieren.

Man kann also nicht generell entscheiden, ob eine große oder eine kleine Lernrate von Vorteil ist. Die Erfahrung hat allerdings gezeigt, daß es sich empfielt, mit einer Lernrate um 0,7 zu beginnen.

Wird nach einer ausreichend langen Trainingszeit keine befriedigende Lerngüte erreicht, empfielt es sich, die Lernrate schrittweise um 0,1 zu verringern.

Bevor man voreilig die Lernrate η verändert, sollte man sich allerdings über die Bedeutung der Initialisierung der Gewichte im Klaren sein.

Initialisierung der Gewichte

Die Initialisierung kann entscheidenen Einfluß auf den Erfolg des Trainings haben, da sie den Startpunkt des Gradientenabstiegs bestimmt. Der Startpunkt entscheidet nicht nur darüber, ob das Verfahren in einem lokalen oder globalen Minimum endet. Startet der Gradientenabstieg in einem kritischen Bereich der Fehleroberfläche, kann es vorkommen, daß das Verfahren gleich zu Beginn oszilliert oder ein langes Plateau bewältigen muß. Der Lernerfolg ist daher eher als mäßig einzustufen.

Aber auch die Art der Initialisierung spielt eine wichtige Rolle. Werden die Gewichte mit zu großen Werten initialisiert, wäre eine noch größere Zerklüftung der Fehleroberfläche zur befürchten, als sowieso schon zu erwarten ist. Werden alle Gewichte gleich initialisiert, führt dies zu dem Problem des Symmetry Breaking.

Bei mangelndem Lernerfolg empfielt es sich somit, zunächst eine andere Initialisierung der Gewichte zu wählen, bevor die Lernrate verändert wird.

Übungsaufgabe

Das folgende Applet stellt eine Weiterführung des Applets am Ende der Herleitung dar. Im Gegensatz zum ersten Applet wird ihnen die Möglichkeit gegeben, alle Parameter des Backpropagation-Netzes selbst zu bestimmen.

Der Farbgebung innerhalb des Applets kommt die selbe Bedeutung zu wie im ersten Applet. Je heller das Grün der Eingabeneuronen und der verborgenen Neuronen ist, umso größer ist die Ausgabe des entsprechenden Neurons. Rote Einfärbung der Ausgabeneuronen deuten einen großen Anteil am Netzwerkfehler an. Ein Übergang von rot nach gelb signalisiert eine Verringerung des Fehlersignals. Positive Gewichte werden blau, negative Gewichte rot eingefärbt.

Um an Informationen über ein spezielles Neuron und ihr Fehlersignal zu gelangen, klicken sie auf das gewünschte Neuron. In dem Textfield werden daraufhin alle Informationen angezeigt und ihre Entwicklung während der Gewichtsmodifikation dokumentiert. Eine Ergänzung zum ersten Applet ist, daß zusätzlich der Bias-Eingang des entsprechenden Neurons angezeigt wird. Weiterhin können Probleme in der Darstellung durch einen Bildschirmrefresh behoben werden.

Sie können alle Gewichte im Netz gleich initialisieren, um das Problem des Symmetry Breaking zu verdeutlichen. Ihnen wird die Möglichkeit gegeben, die Lernrate und die Anzahl der Trainingszyklen den Gegebenheiten der Fehleroberfläche anzupassen. Die Gewichte können ONLINE oder OFFLINE geändert werden und natürlich ist die Anzahl der Neuronen in der verborgenen Schicht frei wählbar. Zudem wird ihnen die Möglichkeit gegeben, die Generalisierungsfähigkeit des durch sie trainierten Netzes abzuschätzen.

Zurück zum letzten Kapitel  Zum nächsten Kapitel