Backpropagation and minimization

1. The backpropagation algorithm

Suppose the input to the neural network is $\mathbf{x}$ and the output is $\mathbf{y}$. The purpose of the backpropaation algorithm is to calculate the derivative of $\mathbf{y}$ with respect to the weights and thresholds of the network. Recall the definition of a feed-forward network:

$$\begin{aligned} \mathbf{b}_0 &= \mathbf{x}, \\ \mathbf{a}_j &= W_j \mathbf{b}_{j-1} - \tau_j, \quad 1 \leq j \leq L, \\ \mathbf{b}_j &= \sigma_j(\mathbf{a}_j), \quad 1 \leq j \leq L, \\ \mathbf{y} &= \mathbf{b}_L. \end{aligned} \tag{23}$$

Usually we take σ_L to be the identity. This definition contains the possibility that each function σ_j consists of a vector of real-valued functions of a real variable, but in general one assumes that all these functions are the same, that is σ_j is applied to each component of the vector $\mathbf{a}_j$. Now we define the functions f_j so that

$$f_j(\mathbf{b}_j) = \mathbf{y}. \tag{24}$$

From the definition (23) we get

$$f_{j-1}(\mathbf{b}_{j-1}) = f_j\big(\sigma_j(W_j \mathbf{b}_{j-1} - \tau_j)\big).$$

Differentiating both sides we get

$$f'_{j-1}(\mathbf{b}_{j-1}) = f'_j(\mathbf{b}_j)\sigma'_j(\mathbf{a}_j)W_j. \tag{25}$$

In this formula $\sigma_j'(\mathbf{a}_j)$ is a diagonal matrix. Since $f_L'(\mathbf{b}_L) = I$ we can use this formula to calculate all the derivatives $f_j'(\mathbf{b}_j)$. Having calculated these we get

$$\frac{\partial \mathbf{y}}{\partial W_j(k,m)} = f_j'(\mathbf{b}_j)\sigma_j(\mathbf{a}_j(k))\mathbf{b}_{j-1}(m), \tag{26}$$

and

$$\frac{\partial \mathbf{y}}{\partial \mathbf{t}au_j(k)} = -f_j'(\mathbf{b}_j)\sigma_j(\mathbf{a}_j(k)). \tag{27}$$

2. The minimization problem

Let $\mathbf{w}$ be a vector that contains all the matrices W_j and threshold-vectors τ_j, $j = 1, \ldots, L$. We denote the output of a neural network with parameters $\mathbf{w}$ and input $\mathbf{x}$ by $\mathbf{y} = \eta(\mathbf{x}, \mathbf{w})$. Suppose that we want the network to be such that $\eta(\mathbf{x}_i, \mathbf{w}) = \mathbf{y}_i$. Then one very reasonable criterion for choosing the parameters $\mathbf{w}$ is to minimize the function

$$E(\mathbf{w}) = \tfrac{1}{2}\sum_{i=1}^{n} |\eta(\mathbf{x}_i, \mathbf{w}) - \mathbf{y}_i|^2.$$

With the aid of the backpropagation algorithm, it is straightforward to calculate the derivative $E'(\mathbf{w})$ and thus a large number of minimiization algorithms can be used.

3. The conjugate gradient method

The conjugate gradient method for finding the the minimum of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as follows:

Definition 15. *If $f \in \mathbb{C}(\mathbb{R}^d)$ is differentiable and $\mathbf{w}_0 \in \mathbb{R}^d$, then $\mathbf{s}_0 = -f'(\mathbf{w}_0)^{\mathrm{T}}$, and if $k \geq 0$ and $\mathbf{w}_k$ and $\mathbf{s}_k$ are determined then $\mathbf{w}_{k+1} = \mathbf{w}_k + t_k\mathbf{s}_k$ where t_k is chosen so that $f(\mathbf{w}_k + t_k\mathbf{s}_k) = \min_{t\in\mathbb{R}} f(\mathbf{w}_k + t\mathbf{s}_k)$, and*

$$\mathbf{s}_{k+1} = -f'(\mathbf{w}_{k+1})^{\mathrm{T}} + \frac{|f'(\mathbf{w}_{k+1})|^2}{|f'(\mathbf{x}_k)|^2}\mathbf{s}_k. \tag{28}$$

In practice the function f is, of course, not a quadratic, at least exactly, but if it is, then the conjugate gradient method works very efficiently.

Theorem 16. *If $f(\mathbf{w}) = \frac{1}{2}\mathbf{w} \cdot A\mathbf{w} + \mathbf{b} \cdot \mathbf{w} + c$, $\mathbf{w} \in \mathbb{R}^d$, where A is a positive definite symmetric $d \times d$ matrix, then the conjugate gradient method terminates in at most d steps.*

Proof. Let us use the notation $\mathbf{g}_k = f'(\mathbf{w}_k)^{\mathrm{T}}$. The method terminates when $\mathbf{g}_k = \mathbf{0}$ and we have to show that this happens when $k \le d$, and for this reason we assume that $\mathbf{g}_k \ne \mathbf{0}$ for $k = 0, \dots, d$.

It follows from the definition of the method that

$$\mathbf{g}_{j+1} \cdot \mathbf{s}_j = 0, \quad j \ge 1. \tag{29}$$

One consequence of this is that since $\mathbf{g}_k \ne \mathbf{0}$ we also have $\mathbf{s}_k \ne \mathbf{0}$ for $k = 0, \dots, d$. Another consequence is that by (28) we have

$$\mathbf{s}_k \cdot \mathbf{g}_k = -|\mathbf{g}_k|^2, \quad k \ge 0. \tag{30}$$

We have to show using induction that the following claims hold for $0 \le j < k \le d$:

$$\mathbf{s}_j \cdot A\mathbf{s}_k = 0, \tag{31}$$

$$\mathbf{g}_j \cdot \mathbf{g}_k = 0. \tag{32}$$

The fact that the function is quadratic implies that

$$\mathbf{g}_{k+1} = \mathbf{g}_k + t_k A\mathbf{s}_k, \quad k \ge 0, \tag{33}$$

where we then by (29) must have

$$t_k = -\frac{\mathbf{g}_k \cdot \mathbf{s}_k}{\mathbf{s}_k \cdot A\mathbf{s}_k}. \tag{34}$$

If $k = 0$ the claims (31) and (32) are empty and therefore hold. Suppose they hold for some $k \ge 0$. Then by (34),

$$\mathbf{g}_j \cdot \mathbf{g}_{k+1} = \mathbf{g}_j \cdot \mathbf{g}_k + t_k \mathbf{g}_j \cdot A\mathbf{s}_k = \mathbf{g}_j \cdot \mathbf{g}_k - t_k(\mathbf{s}_j - \beta_{j-1}\mathbf{s}_{j-1}) \cdot A\mathbf{s}_k,$$

where $\beta_j = \frac{|\mathbf{g}_{j+1}|^2}{|\mathbf{g}_j|^2}$. From this equation one sees that $\mathbf{g}_k \cdot \mathbf{g}_{k+1} = 0$ by (31), (30), and (35). If $j < k$ then $\mathbf{g}_j \cdot \mathbf{g}_{k+1} = 0$ by (31) and (32).

Furthermore we have by (28) and (34)

$$\begin{aligned} \mathbf{s}_j \cdot A\mathbf{s}_{k+1} = -\mathbf{s}_j \cdot A\mathbf{g}_{k+1} + \beta_k \mathbf{s}_j \cdot A\mathbf{s}_k &= -A\mathbf{s}_j \cdot \mathbf{g}_{k+1} + \beta_k \mathbf{s}_j \cdot A\mathbf{s}_k \\ &= \frac{1}{t_j}(\mathbf{g}_j - \mathbf{g}_{j+1}) \cdot \mathbf{g}_{k+1} + \beta_k \mathbf{s}_j \cdot A\mathbf{s}_k. \end{aligned}$$

From this we conclude that $\mathbf{s}_k \cdot A\mathbf{s}_{k+1} = 0$ by (32), (30), (35), and by the definition of β_k. If $j < k$ we have $\mathbf{s}_j \cdot A\mathbf{s}_{k+1} = 0$ by (31) and (32).

If none of the vectors $\mathbf{g}_k$ is zero for $k = 0, \dots, d$ we have found $d + 1$ orthogonal nonzero vectors in $\mathbb{R}^d$ which is impossible. This contradiction completes the proof. □

We have the following partial result on the convergence of the conjugate gradient method.

Theorem 17. *Assume that* $\mathbf{w}_0 \in \mathbb{R}^d$ *and that* $f \in \mathcal{C}^2(\mathbb{R}^d)$ *are such that the set* $\{\, \mathbf{w} \in \mathbb{R}^d \mid f(\mathbf{w}) \le f(\mathbf{w}_0) \,\}$ *is bounded. Let* $\mathbf{s}_0 = \mathbf{0}$ *and suppose that the sequences* $(\mathbf{w}_k)_{n=0}^{\infty}$, $(\mathbf{s}_k)_{n=0}^{\infty}$ *and* $(t_k)_{n=0}^{\infty}$ *are such that for each* $k \ge 0$,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + t_k \mathbf{s}_k, \tag{35}$$

$$\left|\mathbf{g}_{k+1} \cdot \mathbf{s}_k\right| \le -\sigma \mathbf{g}_k \cdot \mathbf{s}_k, \tag{36}$$

$$f(\mathbf{w}_{k+1}) \le f(\mathbf{w}_k) + \rho t_k \mathbf{g}_k \cdot \mathbf{s}_k, \quad t_k > 0 \tag{37}$$

$$\mathbf{s}_{k+1} = -g_{k+1} + \beta_k \mathbf{s}_k, \tag{38}$$

$$\beta_k = \frac{|\mathbf{g}_{k+1}|^2}{|\mathbf{g}_k|^2}, \tag{39}$$

where $\mathbf{g}_k = f'(\mathbf{w}_k)^{\mathrm{T}}$, $\rho > 0$, *and* $0 < \sigma < \frac{1}{2}$.
Then $\liminf_{k\to\infty} \mathbf{g}_k = \mathbf{0}$.

Proof. We leave it as an exercise to show that

$$-\frac{1}{1-\sigma} \le \frac{\mathbf{g}_k \cdot \mathbf{s}_k}{|\mathbf{g}_k|^2} \le -\frac{1-2\sigma}{1-\sigma}. \tag{40}$$

Since we assume that $\sigma < \frac{1}{2}$ it follows that we actually have $\mathbf{g}_k \cdot \mathbf{s}_k < 0$ for all k such that $\mathbf{g}_k \ne 0$.

Furthermore, we have by (37) and (41) that

$$|\mathbf{g}_{k+1} \cdot \mathbf{s}_k| \le -\sigma \mathbf{g}_k \cdot \mathbf{s}_k \le \frac{\sigma}{1-\sigma} |\mathbf{g}_k|^2.$$

Using this inequality together with the definitions (39) and (40) we get

$$|\mathbf{s}_{k+1}|^2 = |\mathbf{g}_{k+1}|^2 - 2\beta_k \mathbf{g}_{k+1} \cdot \mathbf{s}_k + \beta_k^2 |\mathbf{s}_k|^2 \le \frac{1+\sigma}{1-\sigma} |\mathbf{g}_{k+1}|^2 + \frac{|\mathbf{g}_{k+1}|^4 |\mathbf{s}_k|^2}{|\mathbf{g}_k|^4}.$$

Using this inequality in the induction step we can show that

$$|\mathbf{s}_k|^2 \le \frac{1+\sigma}{1-\sigma} |\mathbf{g}_k|^4 \sum_{j=0}^{k} |\mathbf{g}_j|^{-2}. \tag{41}$$

If $\liminf_{k\to\infty} \mathbf{g}_k \ne \mathbf{0}$ (which includes the assumption that $\mathbf{g}_k \ne 0$ for all k) then there is a constant $\epsilon > 0$ such that

$$|\mathbf{g}_k| \ge \epsilon > 0, \tag{42}$$

for all k. Since the points $\mathbf{w}_k$ are contained in abounded set the numbers $|\mathbf{g}_k|$ are bounded from above and there is by (42) a constant c_1 such that

$$|\mathbf{s}_k|^2 \le c_1(k+1), \quad k \ge 0. \tag{43}$$

Using inequality (41) we get

$$\sum_{k=0}^{n} \frac{|\mathbf{g}_k \cdot \mathbf{s}_k|^2}{|\mathbf{s}_k|^2} \ge \left(\frac{1-2\sigma}{1-\sigma}\right)^2 \sum_{k=0}^{n} \frac{|\mathbf{g}_k|^4}{|\mathbf{s}_k|^2} \ge \left(\frac{1-2\sigma}{1-\sigma}\right)^2 \frac{\epsilon^4}{c_1} \sum_{k=0}^{n} \frac{1}{k+1}, \tag{44}$$

and this goes to infinity as $n \to \infty$.

It follows from our assumptions that f'' is bounded on compact sets and therefore there is a constant c_2 such that

$$|\mathbf{g}_{k+1} \cdot \mathbf{s}_k - \mathbf{g}_k \cdot \mathbf{s}_k| \leq c_2 |\mathbf{w}_{k+1} - \mathbf{w}_k| |\mathbf{s}_k|.$$

A consequence of this is that

$$t_k \geq -\frac{1-\sigma}{c_2} \frac{\mathbf{g}_k \cdot \mathbf{s}_k}{|\mathbf{s}_k|^2},$$

and by (38) this implies that

$$f(\mathbf{w}_{k+1}) \leq f(\mathbf{w}_k) - \frac{\rho(1-\sigma)}{c_2} \frac{|\mathbf{g}_k \cdot \mathbf{s}_k|^2}{|\mathbf{s}_k|^2}.$$

But since $f(\mathbf{w}_k)$ is bounded from below this inequality implies that $\sum_{k=0}^{n} \frac{|\mathbf{g}_k \cdot \mathbf{s}_k|^2}{|\mathbf{s}_k|^2}$ is bounded from above, which is a contradiction in view of (45). This completes the proof. □