HUT , Institute of mathematics Gripenberg
Mat-1.196 Mathematics of neural networks
Exercise 9
12.3–21.3.2002

1. Assume that $\mathbf{x} \in \mathbb{R}^d$. Calculate the gradient of the function

$$f(\mathbf{w}) = \tfrac{1}{2}|\mathbf{x} - (\mathbf{w}\cdot\mathbf{x})\mathbf{w}|^2, \quad \mathbf{w} \in \mathbb{R}^d,$$

at a point where $|\mathbf{w}| = 1$.

Solution: We have

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{x}\|^2 - (\mathbf{w}\cdot\mathbf{x})^2 + \tfrac{1}{2}(\mathbf{w}\cdot\mathbf{x})^2\|\mathbf{w}\|^2.$$

If $1 \le j \le d$ then a straightforward calculation shows that

$$\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}(j)} = -2(\mathbf{w}\cdot\mathbf{x})\mathbf{x}(j) + (\mathbf{w}\cdot\mathbf{x})\mathbf{x}(j)\|\mathbf{w}\|^2 + (\mathbf{w}\cdot\mathbf{x})^2\mathbf{w}(j) = -(\mathbf{w}\cdot\mathbf{x})\mathbf{x}(j) + (\mathbf{w}\cdot\mathbf{x})^2\mathbf{w}(j),$$

where we used the fact that $|\mathbf{w}| = 1$. Thus the gradient of f is

$$f'(\mathbf{w}) = -(\mathbf{w}\cdot\mathbf{x})\mathbf{x} + (\mathbf{w}\cdot\mathbf{x})^2\mathbf{w}.$$

Note that if we want to minimize f for a given fixed $\mathbf{x}$ we choose, of course, $\mathbf{w} = \frac{1}{|\mathbf{x}|}\mathbf{x}$, but if $\mathbf{x}$ is not fixed but a random vector, this is not a good idea. However, if we are given a sequence of samples of this random vector, we can at each step change $\mathbf{w}$ a small step in the direction of the negative gradient, and then we get the updating formula

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \gamma_n((\mathbf{w}_n\cdot\mathbf{x}_n)\mathbf{x}_n - (\mathbf{w}_n\cdot\mathbf{x}_n)^2\mathbf{w}_n).$$

---

2. Assume that $\mathbf{w} \in \mathbb{R}^d$ is such that $|\mathbf{w}| = 1$. If $\mathbf{x} \in \mathbb{R}^d$, define

$$\mathbf{w}^* = \frac{1}{|\mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})\mathbf{x}|}(\mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})\mathbf{x}),$$

so that $|\mathbf{w}^*| = 1$. Show that

$$\mathbf{w}^* = \mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})(\mathbf{x} - (\mathbf{w}\cdot\mathbf{x})\mathbf{w}) + O(\eta^2).$$

Solution: Because $|\mathbf{w}| = 1$ and since $\frac{1}{\sqrt{1+t}} = 1 - \frac{1}{2}t + O(t^2)$ we have

$$\frac{1}{|\mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})\mathbf{x}|} = \frac{1}{\sqrt{|\mathbf{w}|^2 + 2\eta(\mathbf{w}\cdot\mathbf{x})^2 + \eta^2(\mathbf{w}\cdot\mathbf{x})^2|\mathbf{x}|^2}} = 1 - \eta(\mathbf{w}\cdot\mathbf{x})^2 + O(\eta^2).$$

Hence

$$\begin{aligned}\frac{1}{|\mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})\mathbf{x}|}(\mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})\mathbf{x}) &= (\mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})\mathbf{x})(1 - \eta(\mathbf{w}\cdot\mathbf{x})^2 + O(\eta^2)) \\ &= \mathbf{w} + \eta(\mathbf{w}\cdot\mathbf{x})(\mathbf{x} - (\mathbf{w}\cdot\mathbf{x})\mathbf{w}) + O(\eta^2).\end{aligned}$$

---

3. Assume that $x \in \mathbb{R}^{d\times 1}$. Find the gradient of the function

$$f(W) = \tfrac{1}{2}|X - W^{\mathrm{T}}WX|^2, \quad W \in \mathbb{R}^{m\times d},$$

at a point where $WW^{\mathrm{T}} = I$.

Solution: We can rewrite $f(W)$ as

$$f(W) = \tfrac{1}{2}\left(X^{\mathrm{T}}X - 2X^{\mathrm{T}}W^{\mathrm{T}}WX + X^{\mathrm{T}}W^{\mathrm{T}}WW^{\mathrm{T}}WX\right).$$

If we have a function of the form $g(W) = AWB$ where A and B are matrices with the dimensions $1 \times m$ and $d \times 1$, then

$$\frac{\partial g(W)}{\partial W(j,k)} = A(1,j)B(k,1),$$

so we can write

$$g'(W) = A^{\mathrm{T}}B^{\mathrm{T}}.$$

Similarly if $g(W) = AW^{\mathrm{T}}B$, then $g(W) = B^{\mathrm{T}}WA^{\mathrm{T}}$ so that $g'(W) = BA$. Applying these differentiation rules to f we get

$$f'(W) = \tfrac{1}{2}\big(-2WXX^{\mathrm{T}} - 2WXX^{\mathrm{T}} + WW^{\mathrm{T}}WXX^{\mathrm{T}} + WXX^{\mathrm{T}}W^{\mathrm{T}}W$$
$$+ WXX^{\mathrm{T}}W^{\mathrm{T}}W + WW^{\mathrm{T}}WXX^{\mathrm{T}}\big) = -WXX^{\mathrm{T}} + WXX^{\mathrm{T}}W^{\mathrm{T}}W,$$

when we use $WW^{\mathrm{T}} = I$.