Machine Intelligence 1 - Exercise 4: Multilayer Perceptrons and Backpropagation Algorithm

Liu, Zhiwei 387571
Moon, Chulhyun 392865
Wenzel, Daniel 365107
Ozmen, Cengizhan 388011
Pipo, Aiko 390011

November 15, 2017

H4.1: Line search (4 points)

(a)

The multi-dimensional Taylor approximation of a fucntion f around a given point a can be expressed using multi-index notation:

$$T_n f(x;a) = \sum_{|\alpha|=0}^{n} \frac{(x-a)^\alpha}{\alpha!} D^\alpha f(a)$$

Applying this to our case and setting $\underline{w}_{t+1} = \underline{w}_t - \eta_t \underline{d}_t$ yields:

$$\begin{aligned} T_2 E^T(\underline{w}_{t+1}; \underline{w}_t) &= E^T(\underline{w}_t) + \sum_{i=1}^{d} (\underline{w}_{t+1\,i} - \underline{w}_{t\,i}) \frac{\delta E^T}{\delta \underline{w}_i} + \sum_{i=1}^{d}\sum_{j=1}^{d} (\underline{w}_{t+1\,i} - \underline{w}_{t\,i})(\underline{w}_{t+1\,j} - \underline{w}_{t\,j}) \frac{\delta E^T}{\delta \underline{w}_i \delta \underline{w}_j}(\underline{w}_t) \\ &= E^T(\underline{w}_t) + \sum_{i=1}^{d} (-\eta_t \underline{d}_{t\,i}) \frac{\delta E^T}{\delta \underline{w}_i} + \sum_{i=1}^{d}\sum_{j=1}^{d} (-\eta_t \underline{d}_{t\,i})(-\eta_t \underline{d}_{t\,j}) \frac{\delta E^T}{\delta \underline{w}_i \delta \underline{w}_j}(\underline{w}_t) \\ &= E^T(\underline{w}_t) - \eta_t \underline{d}_t^T(\nabla E^T(\underline{w}_t)) + \eta_t^2 \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t \end{aligned}$$

(b)

Using the inequation and $E^T(\underline{w}_{t+1}) \approx T_2 E^T(\underline{w}_{t+1}; \underline{w}_t)$, we find:

$$\begin{aligned} E^T(\underline{w}_{t+1}) &\le E^T(\underline{w}_t) \\ E^T(\underline{w}_t) - \eta_t \underline{d}_t^T(\nabla E^T(\underline{w}_t)) + \eta_t^2 \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t &\le E^T(\underline{w}_t) \\ -\eta_t \underline{d}_t^T(\nabla E^T(\underline{w}_t)) + \eta_t^2 \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t &\le 0 \\ \eta_t^2 \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t &\le \eta_t \underline{d}_t^T(\nabla E^T(\underline{w}_t)) \end{aligned}$$

For $\underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t > 0$ and $\underline{d}_t^T(\nabla E^T(\underline{w}_t)) < 0$, we can only choose $\eta_t = 0$. In all other cases we can simplify the inequation:

$$\eta_t \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t \le \underline{d}_t^T(\nabla E^T(\underline{w}_t))$$

We obtain the following cases (for simplicity $H = \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t$ and $E = \underline{d}_t^T(\nabla E^T(\underline{w}_t))$):

- $\eta_t \le \frac{E}{H}$, if $H, E < 0$ or $H, E > 0$
- $\eta_t > 0$ arbitrary, if $H < 0, E > 0$

(c)

With this cost function we have

$$T_2E^T(\underline{w}_{t+1};\underline{w}_t) = \frac{1}{2}(w_t - w^*)^T H(w_t - w^*) - \eta_t \underline{d}_t^T(\nabla E^T(\underline{w}_t)) + \eta_t^2 \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t$$

According to chain rule,

$$\frac{\partial E^T \underline{w}_{t+1}}{\partial \eta} = \frac{\partial E^T \underline{w}_{t+1}}{\partial w_{t+1}} \frac{\partial w_{t+1}}{\partial \eta}$$

since:

$$w_{t+1} = w_t - \eta d_t$$

we can get:

$$\begin{aligned}\frac{\partial E^T \underline{w}_{t+1}}{\partial \eta} &= \frac{\partial \frac{1}{2}(w_t - w^*)^T H(w_t - w^*) - \eta_t \underline{d}_t^T(\nabla E^T(\underline{w}_t)) + \eta_t^2 \underline{d}_t^T \underline{H}(\underline{w}_t)\underline{d}_t}{\partial w_{t+1}} \frac{\partial (w_t - \eta d_t)}{\partial \eta} \\ &= (\frac{1}{2}H(w_t - w^*) - \eta d_t^T H d_{t)}) \times (-d_t)\end{aligned}$$

make it zero, then:

$$\eta^* =$$